STATUS OF THESIS

Title of thesis: A DECISION SUPPORT SYSTEM FRAMEWORK FOR SEASONAL ZOONOSIS PREDICTION

I ADHISTYA ERNA PERMANASARI ,

hereby allow my thesis to be placed at the Information Resource Center (IRC) of Universiti Teknologi PETRONAS (UTP) with the following conditions:

1. The thesis becomes the properties of UTP.
2. The IRC of UTP may make copies of the thesis for academic purposes only.
3. This thesis is classified as

□ Confidential

☑ Non-confidential

If this thesis is confidential, please state the reason:
-

The contents of the thesis will remain confidential for - years.

Remarks on disclosure:
-

Endorsed by

Signature of Author	Signature of Supervisor
Permanent Address:	Name of Supervisor:
Perum. Dayu Permai B-42	Dr. Dayang Rohaya Awang Rambli
Yogyakarta, Indonesia	
Date:	Date:

UNIVERSITI TEKNOLOGI PETRONAS

DISSERTATION TITLE: A DECISION SUPPORT SYSTEM FRAMEWORK
FOR SEASONAL ZOONOSIS PREDICTION

by

ADHISTYA ERNA PERMANASARI

The undersigned certify that they have read, and recommend to the Postgraduate Studies Programme for acceptance this thesis for the fulfillment of the requirements for the degree stated.

Signature: ______________________________________

Main supervisor: Dr. Dayang Rohaya Awang Rambli_______

Signature: _____________________________________

Co-Supervisor : Assoc. Prof. Dr. P. Dhanapal Durai Dominic

Signature: _____________________________________

Head of Department: Dr. Mohd. Fadzil Bin Hassan____________

Date: _____________________________________

A DECISION SUPPORT SYSTEM FRAMEWORK FOR SEASONAL ZOONOSIS PREDICTION

by

ADHISTYA ERNA PERMANASARI

A Thesis

Submitted to the Postgraduate Studies Programme

as a Requirement for the Degree of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER AND INFORMATION SCIENCE

UNIVERSITI TEKNOLOGI PETRONAS

BANDAR SERI ISKANDAR,

PERAK

SEPTEMBER 2010

DECLARATION OF THESIS

Title of thesis: A DECISION SUPPORT SYSTEM FRAMEWORK FOR SEASONAL ZOONOSIS PREDICTION

I ADHISTYA ERNA PERMANASARI

hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTP or other institutions.

	Witnessed by
______________________________	______________________________
Signature of Author	Signature of Supervisor
Permanent Address:	Name of Supervisor:
Perum. Dayu Permai B-42	Dr. Dayang Rohaya Awang Rambli
Yogyakarta, Indonesia	
Date: __________________________	Date: __________________________

ACKNOWLEDGEMENTS

First, I would like to thank to Allah who made all things good. I gratefully acknowledges to the following people for helping me during my master program. Dr. Dayang Rohaya Awang Rambli, my supervisor, for her assistance and encouragement in the completion of my thesis. This thesis would not be possible without her guidance.

I am thankful to my co-supervisor Assoc. Prof. Dr. P. Dhanapal Durai Dominic, who really helps me in the beginning of this works. He was always available and helpful whenever I had question or problem.

I am also thankful to Prof. Dr. PA Venkatachalam, who really helps me in the beginning of this works and shares his idea with me.

I give a great deal of thanks to Prof. Dr. Rose Alinda Bt Alias and Assoc. Prof. Dr. Abas bin Md Said as examiner members for all feedback and suggestion in improving my research.

I am grateful to Assoc. Prof. Dr. Mohd Noh B. Karsiti, Director of Postgraduate Studies and Dr. Mohd Fadzil Hassan, Head of Computer and Information Science Department, for giving me resources to make this thesis completed.

I would like also express my gratitude to Samkhan DVM., M.Vet.Sc. as a Head of Veterinary Information Section at Disease Investigation Center Yogyakarta, Indonesia, for his discussion relate to this area during his time.

Special thanks to all my colleagues in Universiti Teknologi PETRONAS and Gadjah Mada University for their support, input and friendship.

I am especially thankful for the considerable support of my family. My parents have always giving me their great love in all my love and encourage me to pursuing my higher education. I am grateful to my parents in law for all their supports. I also want to thank to my sisters for their prayer.

I am indebted to my husband and my son. My husband Rahmat Riza has always support me in finishing my PhD. He gives all his love and always beside me whenever I want and my son Athaya Muhammad Daffa Riza is always makes my daily activities more cheerful.

Last but not least, I am thankful to University Technologi PETRONAS for providing grant and facilities for the research.

ABSTRACT

The arising number of zoonosis epidemics and the potential threat to human highlight the need to apply stringent system to contend zoonosis outbreak. Zoonosis is any infectious disease that is able to be transmitted from other animals, both wild and domestic, to humans. The increasing number of zoonotic diseases coupled with the frequency of occurrences, especially lately, has made the need to study and develop a framework to predict future number of zoonosis incidence. Unfortunately, study of literatures showed most prediction models are case-specific and often based on a single forecasting technique.

This research analyses and presents the application of a decision support system (DSS) that applied multi forecasting methods to support and provide prediction on the number of zoonosis human incidence. The focus of this research is to identify and to design a DSS framework on zoonosis that is able to handle two seasonal time series type, namely additive seasonal model and multiplicative seasonal model. The first dataset describes the seasonal data pattern that exhibited the constant variation, while the second dataset showed the upward/downward trend. Two case studies were selected to evaluate the proposed framework: Salmonellosis and Tuberculosis for additive time series and Tuberculosis for multiplicative time series. Data was collected from the number of human Salmonellosis and Tuberculosis incidence in the United States published by Centers for Disease Control and Prevention (CDC). These data were selected based on availability and completeness.

The proposed framework consists of three components: database management subsystem, model management subsystem, and dialog generation and management subsystem. A set of 168 monthly data (1993–2006) of Salmonellosis and Tuberculosis was used for developing the database management subsystem. Six forecasting methods, including five statistical methods and one soft computing method, were applied in the model management subsystem. They were regression analysis, moving average, decomposition, Holt-Winter’s, ARIMA, and neural network. The results of each method were compared using ANOVA, while Duncan Multiple Range Test was employed to identify the compatibility of each method to the time series. Coefficient of Variation (CV) was used to determine the most appropriate method among them. In the user interface subsystem, “What If” (sensitivity) analysis was chosen to construct this component. This analysis provided the fluctuation of forecasting results which was influenced by the changes in data. The sensitivity analysis was able to determine method with the highest fluctuation based on data update. Observation of the result showed that regression analysis was the fittest method for Salmonellosis and neural network was the fittest method of Tuberculosis. Thus, it could be concluded that results difference of both cases was affected by the available data series. Finally, the design of Graphical User Interface (GUI) was presented to show the connectivity flow between all DSS components.

The research resulted in the development of a DSS theoretical framework for a zoonosis prediction system. The results are also expected to serve as a guide for further research and development of DSS for other zoonosis, not only for seasonal zoonosis but also for nonseasonal zoonosis.

ABSTRAK

Peningkatan bilangan wabak zoonosis dan potensi ancamannya kepada manusia telah menimbulkan satu kesedaran untuk membendung penularannya dari berleluasa. Zoonosis adalah penyakit berjangkit yang boleh dipindahkan dari haiwan, liar mahupun domestik kepada manusia. Kenaikan bilangan penyakit zoonotic dan ditambah pula dengan kekerapan kejadian nya lewat waktu ini menimbulkan kesedaran untuk membangunkan dan mengkaji rangka kerja untuk mengjangka kebarangkalian kewujudannya pada masa akan datang. Kajian yang telah diperkenalkan oleh penyelidik-penyelidik awal hanya melibatkan jangkaan berdasarkan model berdasar kes dan juga teknik jangkaan keatas kes tertentu sahaja.

Melalui analisis yang dijalankan dalam kajian ini, aplikasi decision support system (DSS) digunakan sebagai kaedah multijangkaan untuk menyokong dan membolehkan kebolehramalan bilangan zoonosis berlaku pada manusia. Oleh itu, fokus utama kajian ini adalah untuk mengenalpasti dan merekabentuk kerangka kerja DSS untuk zoonosis yang boleh menangani dua jenis siri masa iaitu additive seasonal model dan multiplicative seasonal model. Set data pertama yang dibangunkan dalam kajian ini menerangkan corak data bermusim yang mewakili pemalar variasi, manakala set data kedua menunjukkan bentuk pembolehubah yang sentiasa berubah. Dari sini, dua kajian kes telah dipilih untuk menilai keberkesanan pendekatan yang diambil: Salmonellosis untuk additive time series, dan Tuberculosis untuk multiplicative time series. Data untuk bilangan kejadian Salmonellosis/Tuberculosis keatas manusia yang berlaku di United States dikumpul daripada Pusat Kawalan Penyakit dan Pencegahan (CDC). Data-data dipilih berdasarkan yang boleh didapati pada waktu kajian dan ianya adalah lengkap.

Pendekatan kerangka kerja yang dibangunkan ini terdiri daripada tiga komponen iaitu subsistem pengurusan pengkalan data, subsistem pengurusan model dan antaramuka pengguna. Satu set 168 data bulanan (1993-2006) untuk Salmonellosis dan Tuberculosis digunakan untuk membangunkan subsistem pengurusan pengkalan data. Bagi subsistem pengurusan model, enam kaedah jangkaan telah diaplikasi termasuklah lima kaedah statistik dan satu kaedah soft-computing. Kaedah jangkaan itu adalah Analisis Regresi, purata pergerakan, pembubaran bentuk, Holt-Winter's, ARIMA dan rangkaian neural. Hasil kajian yang diperolehi dari setiap kaedah akan dibandingkan dengan ANOVA, sementara itu Ujian Pelbagai-bidang Duncan digunakan untuk mengenalpasti kebolehsuaian setiap kaedah dengan siri masa. Coeffection of Variation (CV) pula digunakan untuk menentukan kaedah yang paling bersesuaian di antara kaedah yang dipilih. Di dalam subsistem antaramuka pengguna, persoalan mengenai "Apa-Jika" (sensitivity) dianalisis dan telah dipilih untuk membangunkan komponen tertentu. Analisis yang diperolehi dari persoalan "Apa-jika" menyediakan perubahan dalam ramalan keputusan yang mana ianya boleh dipengaruhi oleh perubahan dalam data. Analisis ini juga mampu untuk menentukan kaedah mana yang mempunyai kadar perubahan tertinggi dalam data terkini. Pemerhatian yang diperolehi daripada kajian yang dilakukan menunjukkan Analisis Regresi adalah kaedah terbaik untuk Salmonellosis manakala kaedah Rangkaian Neural adalah yang terbaik untuk Tuberculosis. Dari sini, boleh disimpulkan bahawa perbezaan keputusan dipengaruhi oleh data yang didapati pada sela masa tertentu.

Hasil kajian yang diperolehi daripada pembangunan teoritikal rangka kerja DSS ini digunakan untuk system ramalan zoonosis. Hasil dari kajian ini juga digunapakai sebagai panduan untuk kajian lanjut bagi pembangunan DSS bagi meramalkan jenis zoonosis lain dan meluas kepada zoonosis tidak bermusim.

TABLE OF CONTENTS

Chapter

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	Actual Difference
ADF	Augmented Dickey-Fuller
AI	Avian Influenza
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
ANOVA	Analysis Of Variance
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
BIC	Bayesian Information Criterion
BJ	Box-Jenkins
BPN	Back Propagation Network
BSM	Basic Structural Model
CDC	Center For Disease Control And Prevention
CL	Cutaneous Leishmaniasis
CMA	Centered Moving Average
CV	Coefficient Of Variation
DBMS	Database Management System
df	Degree Of Freedom
DGMS	Dialog Generation And Management System
DHF	Hemorrhagic Dengue Fever
DSS	Decision Support System
EID	Emerging Infectious Diseases
ES	Expert System
FNN	Feed-Forward Neural
GAM	Generalized Additive Models
GDSS	Group DSS
GIS	Geographic Information System

GUI	Graphical User Interface
HIV	Human Immunodeficiency Virus
HPS	Hantavirus Pulmonary Syndrome
KBMS	Knowledge-Based Management Subsystem
LM	Lagrange Multiplier
LSR	Least Significant Range
LSR	Least Significant Range
MA	Moving Average
MAPE	Mean Absolute Percentage Error
MBMS	Model Base Management System
MLP	Multi Layer Perceptron
MMWR	Morbidity And Mortality Weekly Report
MSW	Mean Square Within
NLF	Non-Linear Forecasting
SAC	Sample Autocorrelations
SAR	Seasonal Autoregressive
SARIMA	Seasonal ARIMA
SARS	Severe Acute Respiratory Syndrome
SMA	Seasonal Moving Average
SPAC	Sample Partial Autocorrelations
SS	Sum Of Square
SSW	Sum Of Square Within
TB	Tuberculosis
vCJD	Variant Creutzfeldt-Jakob Disease
WHO	World Health Organization
WNV	West Nile Virus

CHAPTER 1
INTRODUCTION

1.1 Research Background

The contact between human and animal can be seen anywhere, including in public area, such as animal displays, petting zoos, animal swap meets, pet stores, zoological institutions, nature parks, circuses, carnivals, farm tours, livestock-birthing exhibits, county or state fairs, schools, and wildlife photo opportunities [1]. This is how many animal diseases that influence human health can be acquired, because we live in the same environment with some animals [2]. For example at a farm, the farmer must have some interaction with their livestock. It was estimated that 75% of emerging disease infections to humans come from animal origin [3-7]. Thus, the monitoring and evaluation of the effect of animal diseases to the environment, especially on human being is of great concern, to prevent the spread of diseases that originate from animals, which has caused great losses of lives.

"Zoonosis" is defined as any infectious diseases that are transmitted from animals to humans [1, 8]. Zoonotic disease can be transmitted to human through many ways. The Center for Disease Control and Prevention (CDC), U.S., describes some transmission media, in which human can be infected by zoonosis. Fecal-oral route is the primary mode of transmission for enteric pathogens. The animal's fur, skin, and saliva can become contaminated with fecal organisms. Thus, when a person touches a pet animal, or is licked by an animal, a transmission occurs. In other situations, transmission has also occurred from fecal contamination of food, including raw milk, sticky food, and environmental surfaces [1]. Inadequate understanding of zoonosis and its transmission increases the impact of zoonosis for human health.

The evolution of zoonosis from its original form could cause newly emerging zoonotic diseases [6]. Indeed, there has been evidence in a report by WHO [6] associating microbiological factors with the agent, the animal hosts/reservoirs and the human victims, which could result in a new variant of pathogen that is capable of jumping the species barrier. For example, Influenza A virus has jumped from wild waterfowl species to domestic farm, farm animal, and humans. Another recent example is the swine flu, where the outbreaks in human have been detected to originate from a new influenza virus in swine. The outbreaks of a new influenza virus of swine origin started in the United States and have spread to other countries. It has made a disastrous impact to human health around the world, where up to May 2010 there were H1N1 reported cases from more than 214 countries [9].

Some new emerging zoonoses (e.g. avian influenza, Ebola, Marburg, Nipah virus, SARS viruses) and re-emerging zoonosis (e.g. cholera, dengue, measles, meningitis, shigellosis, and yellow fever) have been identified in the 21st century. Besides, the release of biological agent (e.g. BSE/vCJD, anthrax) should be of great concern [6]. Nowadays, some zoonotic diseases can cause a major outbreak in the world. Many people have been killed by zoonotic diseases and many more could be a victim if no further actions are taken by relevant institution.

The statistics by WHO [10] reported some zoonosis outbreaks including Dengue/hemorrhagic dengue fever in Brazil (647 cases, with 48 deaths); Avian Influenza outbreaks in 15 countries (438 cases, 262 deaths) ; Rift Valley Fever in Sudan (698 cases, including 222 deaths); Ebola in Uganda (75 patients), Ebola in Philippines (6 positive cases from 141 suspects) and Ebola in Congo (32 cases, 15 deaths); and the latest was Swine Flu (H1N1) in many countries (over 209,438 cases, at least 2,185 deaths). Some of these zoonoses recently have major outbreaks worldwide which resulted in many losses of lives, both to humans and animals.

Worldwide frequency of zoonosis outbreak in the past 30 years [11] and the risk factor of the newly emerging diseases have forced many governments to apply stringent measures to prevent the outbreak [12], to the extent of destroying livestock in the infected areas. This means great losses to farmers. The impact of zoonosis on human lives shows the need for a modeling approach that allows decision makers to make early estimates of future zoonosis incidences, based on historical time series data.

Predicting future zoonosis incidence is important and will be beneficial in the planning and management of a suitable policy to reduce the number of cases. Generally, information systems play a central role in the development of an effective and comprehensive approach to prevent, detect, respond, and manage infectious disease outbreaks in human [13]. Some works related to management of zoonosis have been done for the past several years [1, 4, 13-27]. The increasing incidence in zoonosis has heightened the need for the development of a prediction system, which is able to provide accurate projection of future incidence. The results can be used for policy/decision makers in developing long term strategies.

Different forecasting methods have been used for the purpose of disease prediction. In some studies, the performance of individual forecasting techniques has been reported, including Multivariate Markov chain model to project the number of tuberculosis (TB) incidence in the United States from 1980 to 2010 [28], exponential smoothing to forecast the number of human incidence of Schistosoma haematobium in Mali [29], ARIMA model to forecast the SARS epidemic in China [30], a Bayesian dynamic model to monitor the influenza surveillance as one factor of SARS epidemic [31], and seasonal autoregressive models to analyze Cutaneous leishmaniasis (CL) incidence in Costa Rica from 1991 to 2001 [32]. It is important to conduct a research that applies the use of various techniques in order to obtain the appropriate techniques that give the more accurate result of prediction. Unfortunately, there are only a few published studies that have compared different forecasting techniques on identical historical data of disease [33-35]. All of these studies only used and compared not more than three techniques and only focus in single zoonosis.

The development of prediction systems in zoonosis has given hope to the prevention and control of outbreak of zoonosis diseases. However, there is a need to develop a general DSS framework on zoonosis to analyze zoonosis trends, and to predict future number of zoonosis incidence. In this research, the framework focused on seasonal zoonosis, because the type of data could be easily collected. Besides, the DSS framework should be applicable for different zoonosis with various patterns, both on the additive and multiplicative seasonal model. The proposed framework used and compared different methods on the same data in order to enable identification of the most accurate prediction among them.

1.2 Problem Statement

The Decision Support System (DSS) has been widely used in various fields. However compared to other areas, such as energy demand prediction, economic field, traffic prediction, and health support, there have been very few studies on zoonosis decision support system. In view of the growing number of emerging zoonotic diseases, it is important to have an early warning system that is able to predict trends of zoonosis. The development of such system is necessary to control the spread of the disease and to protect the public [36].

A number of researches on zoonosis have been conducted, some of which have resulted in the development of zoonosis prediction systems. Many zoonosis prediction studies have been focused on one specific disease only, and most of them use a single forecasting technique in developing the prediction model [14, 17-18, 20, 24-25, 37-45]. Research literatures showed no in-depth study that provided the zoonosis prediction model for multi diseases and using multi forecasting methods. Most existing zoonosis prediction models are case-specific that are not general enough to be used for other diseases. Whilst some have employed several forecasting models, earlier approaches to forecasting is often based on a single forecasting technique. Due to numerous numbers of available forecasting techniques with different performance, selection of the most appropriate prediction model to yield better prediction results becomes critical. Besides, a single technique may not yield the same prediction accuracy for different type of datasets. Thus, this further highlights the importance of selecting an appropriate forecasting model for each specific dataset.

In this research, a DSS framework is proposed to address the problems. The framework is divided into three subsystems, namely database management subsystem, model base management subsystem, and dialog generation management subsystem. Different zoonosis dataset can be involved within the database management subsystem. The model management subsystem applies more than one forecasting technique to predict zoonosis incidence. This subsystem also provides the statistical analysis to find the appropriate technique among them. The framework is capable of not only to provide prediction, but also provide a dialog generation and management component. Using this component, users can access into the system to obtain the prediction result. Users can add some latest data and the system will reprocess the data to produce new prediction based on the latest dataset. Finally, the proposed framework serves as a general guidance that can be applied into others seasonal zoonosis.

1.3 Research Questions

This research aims to provide a general DSS framework to predict future incidence on seasonal zoonosis. To develop the framework, the following research questions are defined:

1. How are current zoonosis prediction system developed in order to support the need of proposing zoonosis prediction framework?

   The sub-questions are:

   - What kinds of zoonotic diseases are being covered?
   - What kinds of techniques are being applied?
   - What kinds of data are being used?

2. What are the components to design the decision support of seasonal zoonosis prediction?

   The sub-questions are:

   - What types of data are being selected?
   - What types of forecasting techniques are being applied?
   - What kinds of sources are being selected to obtain data collection?

- What kinds of user interface are being used?

3. What results are being obtained by applying the real cases study into the proposed framework?

   The sub-questions are:

   - What are the results from additive seasonal model?

   - What are the results from multiplicative seasonal model?

4. How the final analyses can be determined from the result of different case studies?

1.4 Research Aims and Objectives

The overall aim of this research is to develop a decision support system (DSS) framework that uses multi forecasting technique to predict seasonal zoonosis incidence of more than one disease. The framework will be used to predict both additive and multiplicative seasonal time series trends using the appropriate models.

The research has been conducted to meet the following objectives, which will provide the answer to the research questions listed in Section 1.4:

1. To acquire information of previous and current research works dealing with zoonosis prediction system and to identify the research gap that need to be addressed by the proposed model.

2. To develop a DSS framework that can be applied to different seasonal zoonosis that covers two types of time series models, additive time series and multiplicative time series.

3. To apply each DSS component into the selected case studies.

4. To analyze the results comparison of both case studies and to develop GUI design based on DSS component.

1.5 Research Methodology

A general methodology has been formed to address the research questions and research objectives. Basically, the research is divided into four stages:

- Stage 1: Background Study

  Stage 1 introduces background of the research. At this stage the research questions were identified and the research objectives were established to answer the question. The scope of the research is also presented. Literature review of previous works in this area has been conducted to identify the problems and research gap of DSS application for zoonosis prediction. This stage is presented in two chapters, Chapter 1 for introduction and Chapter 2 for literature review.

- Stage 2: Model Design

  This stage is the main part of the research. The research methodology was formulated in this part, and it has been used as the direction of research from the beginning to the final. At this stage, the DSS framework is presented. The framework consists of three DSS components, namely database management subsystem, model base management subsystem, and dialog generation management subsystem. The framework shows the flow of DSS process within each DSS component. Hence, Stage 2 forms the basis and guideline for application of the proposed framework into case studies, which is reported in Stage 3.

- Stage 3: Model Development

  The model development stage involves the application of the research methodology developed at stage 2 into case studies. Stage 3 is divided into two parts based case studies, and is presented in Chapter 4 for Salmonellosis results and Chapter 5 for Tuberculosis results.

- Stage 4: Model Analysis

  This stage discusses and compares the results obtained from the model development. The discussions and comparisons are included in Chapter 6 and the research findings are presented in Chapter 7.

A detailed research methodology is presented in Chapter 3.

1.6 Scope of Research

This research provides a general DSS framework for zoonosis application which provides information and prediction of future incidence of zoonosis in human based on historical data. The DSS framework is focused on zoonosis incidence in human even though zoonosis incidence can be found both in human and animal. Hence, the purpose of the framework is to be able to predict the number of various zoonosis incidences, and not limited to a specific disease only.

Despite the various types of zoonotic diseases, the selected zoonotic diseases are primarily seasonal zoonosis that affect human. The data were collected from annual reports on outbreak of disease from a relevant institution. The data were retrieved and organized on a monthly basis. This historical dataset was used to develop the DSS and predict future number of zoonosis incidence on human. In this research, two different zoonosis dataset were selected: Salmonellosis and Tuberculosis. Salmonellosis was used to represent the application of DSS on additive seasonal model and Tuberculosis for multiplicative seasonal model. These data were selected because of the availability and the completeness of data from Morbidity and Mortality Weekly Report (MMWR) from Centers for Disease Control and Prevention (CDC) US from 1993 to 2007. In addition, the data could be easily recorded from CDC website.

The DSS framework uses six different forecasting techniques as the core of model base management subsystem. Analysis of Variance (ANOVA) and Duncan Multiple are conducted to identify the suitable technique between them. Once the suitable methods are found, Coefficient of Variation (CV) is applied to determine the fittest method. In the generation and management subsystem, What-If Analysis can be used to analyze and measure the fluctuation on forecasting accuracy when added by latest data.

1.7 Research Contributions

This research provides the following contributions as follows:

1. This research provides a review of various prediction model in zoonosis area.
2. This research contributes a general zoonosis DSS framework that can be applied in various seasonal zoonosis. The framework consists of three DSS components: database management subsystem, model base management subsystem, and dialog generation and management subsystem. The DSS has the following features:
   a. The proposed framework addresses a DSS model for two type of time series model: additive seasonal model and multiplicative seasonal model.
   b. In model management component, six different forecasting methods are being applied, which allows for quantitative analysis to select the most suitable method among them.
   c. The framework includes a user interface that is capable of identifying and updating results as new data are being added into the system database.
3. This research provides results of evaluation and analysis of real case studies on the proposed framework.
4. The framework can be used be used as guidance for further study on different zoonosis, either seasonal or non-seasonal zoonosis.

1.8 Organization of Thesis

This thesis is divided into 7 chapters.

Chapter 1 provides a research overview. It describes the research background, research aim and objectives, research contributions, and outlines the overall chapters of the dissertation.

Chapter 2 comprises literature reviews on the basic knowledge in this area. This chapter introduces zoonosis and its impact on human. It also introduces the concept of DSS and forecasting theories that are being used in model management. Following that, some reviews on related works in zoonosis are presented. At the end of Chapter 2, some findings on the related works are reported.

Chapter 3 gives the detailed research methodology for developing the DSS framework. The methodology is divided into four conceptual stages: Stage 1: Research Background, Stage 2: Framework Design, Stage 3: Model Development; and Stage 4: Evaluation. A complete description of each stage is presented to ensure correct path for the model development. The two selected case studies are defined in this chapter i.e. Salmonellosis (additive time series) and Tuberculosis (multiplicative time series). The DSS framework is provided within this chapter to be applied into case studies. An illustration of chapter mapping from Chapter 1 through Chapter 7 based on the four proposed stages is presented at the end of Chapter 3.

Chapter 4 reports the statistical analysis of DSS model in Salmonellosis. It includes results of database management subsystem, model base management subsystem, and dialog generation and management subsystem. The procedure and technique for each component are described clearly.

Chapter 5 presents the result of DSS model in Tuberculosis. Overall, the flow of this chapter is similar to Chapter 4. The difference between Chapter 4 and Chapter 5 is in the forecasting approach, as described in Chapter 3.

Chapter 6 compares the results from cases study on Chapter 4 and Chapter 5. It identifies the similarities and the differences between case studies. Thus, this chapter also presents the causal factor associated with the differences.

Chapter 7 summarizes the content of the dissertation. It highlights the research and emphasizes on the advantages of the proposed DSS framework. The limitation of this work and recommended direction for future research are presented at the end of this chapter. The setting of thesis chapter is illustrated in Figure 1.1.

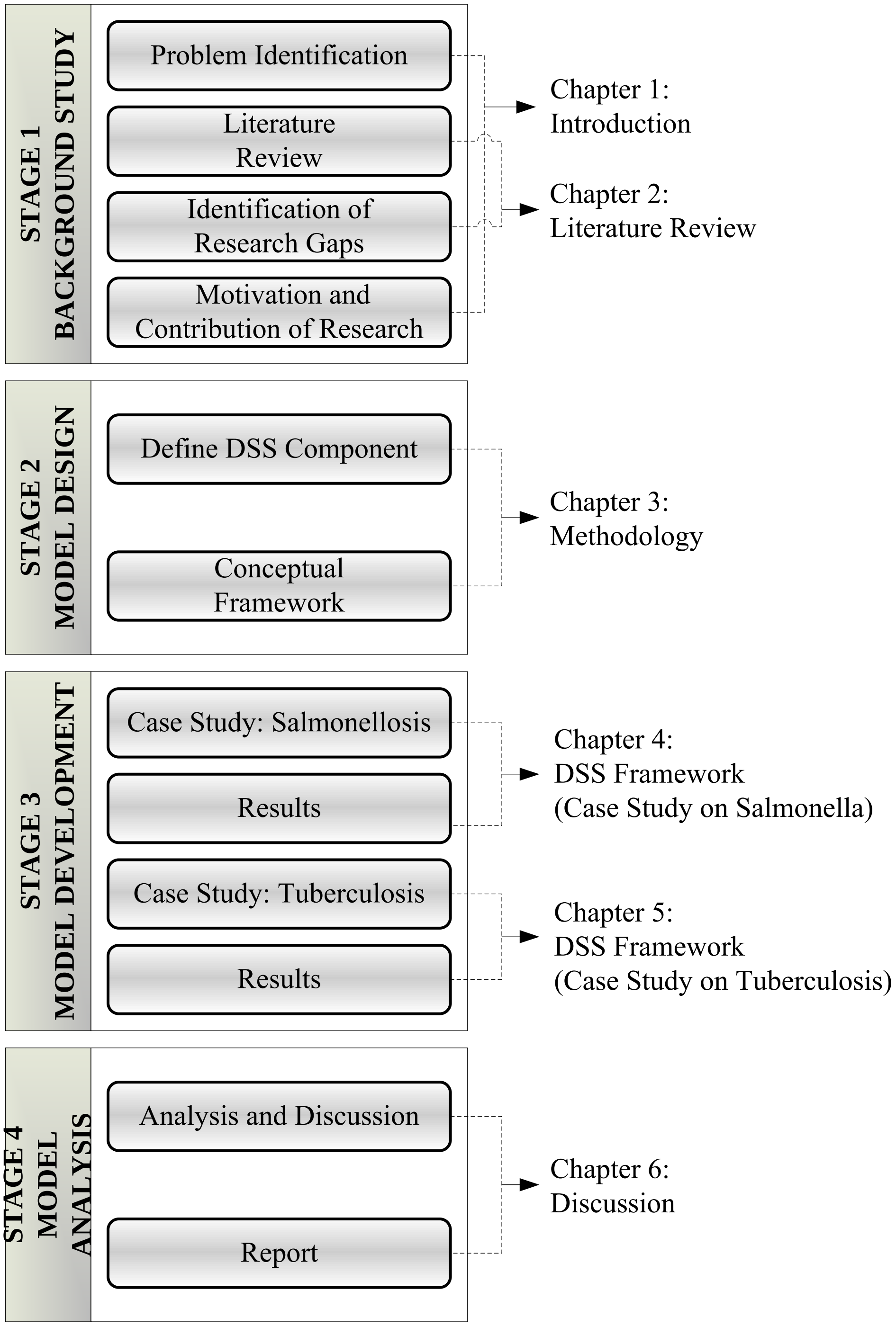


Figure 1.1 Chapter Setting

CHAPTER 2
LITERATURE REVIEW

2.0 Chapter Overview

This research is concerned with the construction of a Decision Support System (DSS) framework in the area of zoonosis. The development of this framework is filled by the wide spectrum of knowledge. In this chapter, a review of the literature research is compiled for providing information to define the area of study, to clarify issues that should be considered, and to support the basic theory of this research.

The chapter begins with Section 2.1 that introduces zoonosis terminology and zoonosis impacts on human being around the world. Section 2.2 defines the basic concepts of DSS. Section 2.3 explores different forecasting theories. Section 2.4 focuses on related works on zoonosis and the use of information systems in zoonosis. The objective of this section is to present the application of different information models to predict zoonosis outbreaks. Section 2.5 summarizes existing systems of zoonosis prediction and highlights some research gaps that need to be addressed for improvement. Finally, Section 2.6 concludes the overall review of Chapter 2.

2.1 Overview of zoonosis

Zoonosis terminology was initially introduced in 1855 by Rudolf Virchow, a German physician, who studied Trichinella [4]. Zoonosis is any infectious disease that is able to be transmitted from animals, both wild and domestic, to humans [1, 4, 7-8, 14, 22, 27, 46-49]. Each zoonosis has different characteristics. Several factors such as zoonosis pathogen, agent, reservoir-host, and mode of transmission to humans give uniqueness to each zoonosis.

A zoonosis pathogen is a biological agent that causes zoonotic disease to a host. Each pathogen can be classified into specific infection agent such as bacteria, parasites, viruses, fungi, and prions [8, 50]. For example Bacillus anthracis is Antrax pathogen and Vibrio cholerae is Cholera pathogen.

A reservoir/host is any species which can transmit zoonosis from its agent and infect another species. A single pathogen disease could have one or more than one reservoir. For example poultry, chicken and pig are reservoir of Avian Influenza.

Humans could be infected by zoonosis from the host through several transmission paths. The contact between human and animal can be seen anywhere, from home, surrounding residence, working place, and public area, such as animal displays, petting zoos, animal swap meets, pet stores, zoological institutions, nature parks, circuses, carnivals, farm tours, livestock-birthing exhibits, county or state fairs, schools, and wildlife photo opportunities [1]. Thus, zoonosis can be transmitted to human through many ways, i.e. direct contact with infected animal, infected food, raw milk, bite/scratch, stitch, and airborne.

Approximately 75% of Emerging Infectious Diseases (EID) comes from animal [3-7]. EID are diseases that have recently increased in incidence, either by geographic location or host range (e.g. tuberculosis, West Nile Fever, and yellow fever). These are caused by new variants of known pathogen (e.g. Avian Influenza virus, SARS, Nipah Virus, and Ebola virus) [2]. A definition for emerging zoonosis was made at the WHO/FAO/OIE joint consultation in Geneva, 3-5 May 2004 [6], which states:

“An emerging zoonosis is a zoonosis that is newly recognized or newly evolved, or that has already occurred previously but shows an increase in incidence or expansion in geographical, host or vector range. Some of these diseases possibly develop and become transmissible between human beings.”

EID can be classified into three categories [27]:

- Newly emerging infections - this group consists of any infectious diseases that have not been previously recognized in human.

Examples are Arenavirus hemorrhagic fevers (Argentine, Bolivian, Venezuelan and Lassa hemorrhagic fevers), hantavirus pulmonary syndrome (HPS), variant Creutzfeldt-Jakob Disease (vCJD), Legionnaire's disease, and Escherichia coli.

- Re-emerging or resurging infections - this is caused by different factors, such as microbial evolutionary vigor, zoonotic encounters and environmental encroachment.

  Examples are the resistance of Plasmodium falciparum, Influenza A virus (Avian Influenza, Swine Flu), SARS, Tuberculosis, Staphylococcus, and Cholera.

- Deliberately emerging infections - this consists of any microbes that have been deliberately developed by human for crime.

  Examples are Anthrax spore sent by terrorist in 2001, attacking 18 people in the US and killing 5 of them, and the trial run of anthrax weapon made by a doomsday sect in Japan in 1995.

Zoonosis evolution from the original form could cause the newly emerging zoonotic disease. It was reported by Dr Cathy Roth [6] that:

> "Epidemic threats continue to be characterized by unexpected events with unstable or poorly understood patterns of transmission and pathogenesis, and bring with them the potential for large public health and economic impact".

For example are the relationship between SARS and Avian Influenza. The research was done by Bush [41] who tried to find the cross transfer mode of unknown virus, which causes SARS, from the previous host to human. Despite the lack of information on SARS evidence, the researchers tried to learn the movement of Avian Influenza virus to human. The research consisted of the three parts. The first part was reanalysis data transmission of Avian Influenza during the Spanish Flu in 1918. The second part was an analysis of viruses archived from three influenza pandemics. The last part was an analysis of limitation of molecular data to study infectious diseases evolution. The final result was able to determine a similar pattern between the diseases and to find any clue about SARS movement.

2.2 Causal Factor of Zoonosis Incidence

There were some factors that took part in zoonosis emergence as stated by Brown [4]. These factors were movement of animals, incultivatable microorganisms, chronic diseases, improved surveillance, terrorism, and ecological discruption.

As stated in Biology Online, the terminology of "*ecology*" has the following meaning:

(1) "Ecology is a branch of biology that deals with the distribution, abundance and interactions of living organisms at the level of communities, populations, and ecosystems, as well as at the global scale."

(2) "Ecology is the system within the environment as it relates to organisms living in it."

Wilcox and Colwell assumed that zoonosis is based on a human-natural system perspective [2]. Hence, it provided a social-ecological approach for gaining a better understanding of EID. A "blueprint" for the proposed interdisciplinary approach was offered there which integrated biological process, incorporating public health infrastructure, and climate aspect. An example of this phenomenon could be seen in the relation between Avian Influenza (H5N1), and the evolutionary and social ecology of infectious disease emergence [51].

The change in social ecology e.g. paradigm shift of livestock husbandry from the family level into industrial scale, increases the opportunity of disease transmission. Ecology could be used as the model-base for defining West Nile Virus (WNV) risk perception [43]. The communication between personal and community took an important part for sharing and designing a framework to reduce the risk of human infected by WNV mosquito-borne virus. A study in relationship between ecotone (the edge or transition zone between two adjacent ecological systems) and disease emergence gave a result that about a half of the approximately 130 zoonotic EID suggested ecotones [52]. The modification of ecotone also could be associated with the global trend of increasing EID.

Slingenbergh et al. [7] argued that changes in agricultural practices became the dominant factor of zoonosis emergence. This factor could determine the conditions of evolution, spread, and entries of zoonosis pathogen evolve, spread, and enter into human population.

The other framework was introduced by Eisenberg *et al.* [53]. The framework examined a relationship between environmental changes and diseases transmission by integrating three interrelated factors of environment-disease relationship: environmental change in ecological and social factor, transmission mode of infectious diseases, and disease burden. The factors were converted into the matrix to define link among them. The framework presented the relationship of environment and disease transmission process that is able to provide further information related to the changes of environment to the infectious disease incidence.

An algorithm for early qualitative public health risk assessment has been developed by Eisenberg *et al.* that exposes the strengths and weaknesses of available evidence on risk. Based on this algorithm, levels of confidence risk of zoonotic transmission can be classified as below [16]:

- Level 0: Not zoonotic — Lack of evidence of zoonotic potential. Good grounds for not taking further action.
- Level 1: Potential zoonosis — Possibility of human pathogenicity not excluded. Work needed on biomarkers of infection and pathways of exposure.
- Level 2: Potential zoonosis — Serological evidence of infection or human exposure has occurred but surveillance is not sufficiently reliable. Enhanced surveillance needed.
- Level 3: Confirmed zoonosis — Human cases have been reported, but evidence against person to person spread. Enhanced surveillance needed. Control exposure of humans to animals and environmental sources.

- Level 4: Confirmed zoonosis — Human cases have occurred, with subsequent person to person spread not excluded. Control of direct or indirect person to person spread needed.

The algorithm was applied to five emerging animal diseases, which are of concern to public health. For example, air borne disease virus can be classified as level 2 zoonosis, bovine norovirus as level 0 zoonosis.

Climatic change has been identified as one of the important factors that causes disease outbreak. Many studies have been carried out to analyze the relationship between climatic change and human health, such as the research by Patz *et al.* [54] who studied the impact of climate fluctuation on human health. Palmer *et al.* [55] reported climatic change of malaria. Recurrent epidemic caused by seasonal dynamic resulted from the work by Stone *et al.* [56]. In this study, they developed a mathematical model that focuses on post-epidemic dynamics. The model was able to identify a new threshold from the measurement of population susceptibility after the last outbreak. It showed that the population susceptibility was influenced by seasonal condition. The evidence obtained from their work became the motivation for other developing forecasting tools associated with a newly emerging and re-emerging disease controlled by seasonal factors.

An insight into emerging zoonotic pathogens as an ecological phenomenon could give an overview as to why some pathogens could jump between different species and causes a number of epidemics in humans. Unfortunately, the complex and non-linear behavior of the environment where host-pathogens are embedded made disease emergence difficult to predict [11].

2.3 Impact of Zoonosis on Human

The number of zoonotic disease incidence and frequency has increased in the past 30 years [11]. Zoonosis occurrences influence public health, social-economic, and political consequences [57]. Large numbers of people have been killed by zoonotic disease in different countries. This problem has forced many governments to apply stringent measures to prevent zoonosis outbreak, for example by destroying the last livestock in the infected area. This means great losses to farmers. Even though zoonotic illnesses also has an economic impact [58], the significant impact to human life, however, is still the biggest issue in zoonosis. Many zoonosis incidences in human have been reported in years, some of which are summarized in Table 2.1.

Table 2.1 Number of Zoonosis Cases

Disease	**Number of cases (in human)**	**Countries**
Swine Flu	More than 622,482 cases, at least 7,826 of them were dead	Worldwide
Avian Influenza	498 cases, 294 deaths (until 6 May 2010)	Azerbaijan, Cambodia, China, Djibouti, Egypt, Indonesia, Iraq, Laos People's Democratic Republic, Nigeria, Thailand, Turkey, Vietnam
Salmonellosis	± 1,400,000 cases, 580 deaths (annually)	USA
Tuberculosis	± 9,369,000 cases, 1,322 deaths	Worldwide
HIV/AIDS	± 40,000,000 cases (the end of 2003)	Worldwide
Nipah Virus	468 cases, 238 deaths (until April 2007)	Worldwide, the biggest victims came from Malaysia
Rabies	50,000 – 100,000 annually	Worldwide, most of them in developing countries
Hantavirus	60,000 – 150,000 annually	90% in Asia (China, Russia, and Korea)
Campylobacter	400 cases per 100,000 people	The highest case number occurs in New Zealand
Monkeypox	37 cases (2003)	USA
E. coli O157:H7	84 cases (2000-2001)	Minnesota, USA

The recent outbreak of Swine influenza originated from swine and has spread around the world. In 2009, this disease became an issue in the world because of its wide spread and was announced as a pandemic by WHO on 11 June 2009 [59]. Firstly, Swine Influenza Virus (SIV) is endemic in pig and the transmission of this virus from pig to human is not common. When the virus transmission causes human influenza, it is called zoonotic swine flu. Among the different influenza A subtypes, H1N1 is the most exclusive strain that attacks human. The first H1N1 pandemic occurred in 1918 [60]. In the early 1976, there was an outbreak that took place in the US army recruit at Fort Dix [61]. In 1988, H1N1 killed a pregnant woman in Wisconsin [62]. However, the biggest and widest H1N1 infection in human started in the US in 2009 and later hundreds of cases occurred in Mexico [63-64]. During that year several outbreaks of H1N1 were reported around the world, with the number of cases in human exceeding 622,482 and at least 7,826 (1.25%) of them were dead [65]. The latest WHO update on 7 May 2010 reported the confirmed H1N1 cases from 214 countries, including mortality number over 18001 [9].

Another recent zoonosis that has claimed many lives is Avian influenza, or “bird flu”, a recent disease of animal origin which has spread widely in Asia [40]. The mortality rate of highly pathogenic avian flu in human is high; WHO data indicated that 60% of cases are classified as Avian; until May 2010 there were 498 cases of AI that attacked human in many countries, with 294 deaths reported [66-67].

Salmonellosis is one of the foodborne diseases. It can be transmitted to human when they consumed contaminated food from animal origin, including meat, poultry, eggs, and milk. Salmonellosis is assumed as a major public health problem in many countries associated to its high cost of impact. Unfortunately, only few countries record the economic cost of diseases. In USA, it is predicted that there is 1.4 million cases annually (with 580 deaths) and it spent the total cost around US$ 3 billion annually [68].

Tuberculosis (TB) is a kind of diseases that is able to be transmitted between human through the air. As predicted by WHO, the largest TB cases occurred in South-East Asia region. In 2008, it was recorded the global number of incidence was

9,369,000 whereas 3,213,000 (34%) cases were from South-East Asia. The number of mortality were 1,322,000 with mortality rate per 100,000 population was 20 [69].

Human Immunodeficiency Virus (HIV)/AIDS become one of the biggest zoonosis pandemic in human. It was estimated by United Nations that 40 million people were infected worldwide by the end of 2003 with more that 3 million deaths reported [22].

In 1999, an outbreak of Nipah Virus was reported in Malaysia. This virus causes inflammation into the brain and respiratory organ. The early case of it was located in Perak, Malaysia, and later in Negeri Sembilan where it destroyed the swine industry in that state [4, 70]. In the beginning, this virus was amplified in the respiratory tracts of swine, and then it was carried in the air to human. The virus goes to the brain, directly. As a result, the infected human would get haemorrhagic and necrotic tracts [4]. In 1999, 265 cases of viral encephalitis in human were recorded in Malaysia, with 38% mortality [22, 70]. Until April 2007, the total number of cases in the world was 468, and from these cases 238 people have died [70].

The other well known zoonosis is rabies. Approximately 50,000 to 100,000 cases of human killed by rabies, a kind of viral disease which transmits to human through infected animal bite, are reported every year. The vast majority of these cases occur in developing countries [22].

Hantavirus is a zoonosis transmitted by rodent. It has been reported that the number of patients is about 60,000 – 150,000 annually, from which more than 90% of these cases occur in Asian countries, including China, Russia, and Korea [42].

The highest rate of campylobacter among countries has occurred in New Zealand, and it has showed a rising number of incidences every year. Helms *et al.* [71] stated that this enteric zoonosis caused an increased risk of mortality based illness. Reportedly, the economic cost from this zoonosis has been estimated approximately between $120 and $220 million [72].

In 2003, several cases of monkey pox were detected in the United States of America among persons who had had contact with infected prairie dogs [1]. It occured in four different states with 37 number of cases in human [4].

A potential consequence of E. coli O157:H7 infection had occurred in Pennsylvania in 2000 involving 59 persons in the Pennsylvania outbreak. From these cases, 51 persons (median age: 4 years) became ill within 10 days after visiting a dairy farm, and eight others (16%) developed hemolytic uremic syndrome (HUS) [1]. Between 2000-2001, 84 cases of E. coli O157:H7 outbreak was reported in Minnesota.

Severe Acute Respiratory Syndrome (SARS) is another dangerous zoonosis, which broke out in some countries, initially found in Guangdong South China in November 2002. The emergence cases in human were identified during 2002-2003 in Southeast Asia [22]. SARS can move into humans through a reservoir species and cause an epidemic in its new host [41].

The great impact of zoonosis to human being can be seen from the statistic given in the table. The zoonosis incidences are not confined in one area, but they may spread to different places. This highlights the need of a prediction system of zoonosis occurrences, which would help to reduce the risk of future incidence.

2.4 Related Work in Zoonosis Prediction Model

Since the increasing number of emerging and re-emerging zoonosis, there is a need on the collaboration between coordinated research, interdisciplinary centre, integrated surveillance systems, response systems and infrastructures, and workforce development strategies [57]. The collaboration could strengthen the partnership in future challenges of emerging zoonosis.

Nowadays, several new emerging zoonoses (swine flu, avian influenza, ebola, marburg, nipah virus, and SARS viruses) and re-emerging zoonosis (cholera, dengue, measles, meningitis, shigellosis, and yellow fever) have been reported. The increasing numbers of recent major zoonoses outbreaks worldwide, including bioterrorist attacks and emerging diseases, recently have major outbreaks worldwide which have resulted in many losses of lives, both humans and animals. This situation calls for the need to develop an early warning system that is able to provide a real-time syndrome surveillance, and an outbreak prediction system [73].

While it is difficult to find previous reviews of DSS specifically focused on zoonosis prediction, there are a number of studies which address closely related themes. The following section describes various researches in zoonosis prediction model. The section is divided into two subsections. The first subsection explores different model on zoonosis prediction based on time series data. In the following section, the use of time series to develop the prediction is added by spatial component and become the temporal-spatial approach.

2.4.1 Zoonosis Prediction Model using Time Series Data

Autoregressive Integrated Moving Average (ARIMA) is the most commonly used method in time series models [74]. This method has the capability to correct the local trend in data, where the pattern in the previous period can be used to forecast the future. Thus this method also supports in modeling one perspective as a function of time (in this case, the number of zoonosis in human case) [75].

Several prediction models have been done based on ARIMA method such as the SARS epidemic in China, which was modeled to monitor the dynamical incidence of the disease [30]. The model provided the results for different purposes, where it consisted of three parts: AR(1) to measure the effectiveness of the SARS interventions; the random walk model to develop a short term forecasting; and the growth curve to model the relative long term effect of SARS. Among the models, the random walk model exhibited the best result and it presented the good fit from the plot. However, the forecast results were influenced by the scale effect, and it caused the variability of forecasted values. Thus, Lai [30] suggested to combine other methods to obtain the better model in analyzing disease dynamic and predict the future incidence.

Shtatland *et al.* [75] reported the potential usefulness of different ARIMA models and logistic regression in biosurveillance. The work was divided into many steps. Within the temporal analysis, they suggested the use of fundamental ARMA models, AR (1), AR(2), and ARMA(1,1), for individual zip code. While, the selection of suitable models was done based on the values of R-squared, Adjusted R-squared, AIC and BIC. Data used in the model development were collected within 30 – 60 days.

Based on the data, the values of R-Squared, adjusted R-Squared and AIC were analyzed. It also resulted in parameters values for each ARIMA model which were determined from the analysis; φ1 for AR(1), (φ1 , φ2) for AR(2), and (φ1,θ1) for ARMA(1,1). This is followed by logistic, Poisson regression or GLMM to narrow the set of surveillance data. Using the study, they showed that two different information between number of patient and geographical location of patient could be combined together to analyze surveillance data.

Box-Jenkins approach was used to apply ARIMA model in predicting the number of dengue incidence in Rio de Janeiro from 1997-2004 [76]. The monthly forecast value could be calculated based on one, two, and twelve prior data. It was used to project the number of incidence in 2005. The process was divided into two alternatives, 12-steps ahead or a one-step ahead. The result showed that the second alternative (one-step ahead) achieved a better accuracy than the first option. An alternative predicting system based on climate was also incorporated into the system. Finally, the model was able to support data surveillance in dengue incidence at that area.

A comparative study was conducted by Sai *et al.* [34] to predict the prevalence of schistosomiasis in Haokou village. The methods used in the study were moving average (MA), exponential smoothing, and ARIMA. Data from 1990-2002 were collected to develop the forecasting model and then, the sum square errors (SSE) of the results were compared to find the fittest model. The result showed ARIMA as the best model among all methods and it has the least SSE value of 26.63.

Earnest *et al.* [77] applied ARIMA to model a real time prediction of number of beds occupied in a hospital during SARS outbreak in Singapore. The data collected from Tan Tock Seng Hospital from 14th March 2003 to 31st May 2003 were divided into two groups, for model development and model testing. The result indicated that ARIMA (1,0,3) was the most appropriate model that achieved accuracy percentage (MAPE) 5.7% for training set and 8.6% for validation. The model also included a three day ahead forecast.

Forecasting of malaria incidence was reported in Karuzi, Burundi by Gomez-Elipe *et al.* [78]. The aim of the research was to develop a forecasting system capable of linking dynamic data and environmental factors. Data of monthly malaria incidence, rainfall, temperature, and vegetation density were recorded in the period 1997-2003. Again in this research, ARIMA methodology was applied and the best forecasting model was selected based on adjusted R^2 values and MAPE. The fitted model was obtained at adjusted $R^2 = 82\%$ and MAPE = 93%.

ARIMA model was also used to forecast Malaria incidence at Ethiopia by Abeku, *et al* [79]. Historical monthly morbidity data from 20 areas in central and north-western Ethiopia were recorded to develop the model. The seasonal variations of data within time interval 1990-1999 were calculated in monthly period. Due to the lognormal distribution of data, the analysis was conducted based on log-transform data series. The most fitted ARIMA model was obtained using Akaike Information Criterion (AIC). The result among all areas yielded the average forecast error up to 9 months period was 0.22. However, this study highlighted the limitations of morbidity pattern in the prediction of malaria incidence where unstable transmissions were detected. It could be improved by adding other factor like meteorological factor.

In the study by Briët *et.* [35], different techniques namely exponentially weighted moving average models, ARIMA models with seasonal components, and multiplicative SARIMA were used in forecasting malaria incidence in Sri Lanka up to four months ahead. The models were developed based on monthly malaria data from 1972-2003. The models were applied in different districts in Sri Lanka. Due to the heterogeneity, every district required different model. Among the methods investigated by Briet *et al.*, ARIMA was able to give the highest accuracy in many districts compared to other techniques. This work also proved that the differences in the time series yielded different results thus require different prediction approaches.

A time series model was applied in Thailand to simulate the effects of climate change on hemorrhagic dengue fever (DHF) [80]. Regression analysis was used in the model with five factors as the equation parameters namely constant, trends, cyclic effects, climatic factors, and noise. Data from 73 provinces were collected to develop the model. The research focused on observing rainfall and temperature level. The results showed that high level of rainfall was not associated with the increasing number of incidence. In contrast, the rise in the temperature had a significant effect to the high number of incidence. Moreover, most incidences were associated with trend and cyclic factors with the incidence variability between 14.7%-75.3%.

In Ethiopia, where malaria epidemic become a great issue and contributes 20% of the annual number of deaths on children under five years, there is a need to predict the number of malaria incidence in order to construct an efficient control decision [81]. Weekly set of malaria data based epidemic in Ethiopia were obtained from 1990 through 2000. Then, the data were normalized to obtain daily mean. Daily meteorological data from the National Meteorological Services Agency (NMSA) in the same period were corresponded into the weekly data. Poisson regression with lagged weather predictor in 4th degree polynomial was used. A modeling system that relates to weather depends on the temperature conditions whether it is hot or cold was developed. The result indicated that the system performed better in cold districts than hot districts.

A research on forecasting of Tuberculosis (TB) incidence in United States was conducted by Debanne *et al.* [28]. The dataset were obtained from the U.S. Bureau of the Census and the Centers for Disease Control and Prevention. The rate of incidence was categorized based on multivariate demography and a fuzzy method was employed in handling parameter uncertainty. The multivariate Markov chain method was used to predict the future number of TB incidence. The model developed was able to give annual projection of TB incidence in the United States from 1980 – 2010 based on race, ethnicity, and geographic groups.

Cutaneous leishmaniasis (CL) is one of the main emergence diseases in America. A study on CL forecasting was conducted by Chaves and Pascual [32]. In the study, a database of monthly number of CL from 1991-2001 was used for model development. The aim of the study was to observe the relationship of CL incidence with the climate conditions. Then, linear model was fitted into CL time series by using climatic predictor to find CL cycle. Climate parameters used in the model were temperature and Multivariate ENSO Index (MEI). The result showed that the linear model could predict dynamical CL incidence up to 12 months ahead. It was determined the variability of the forecast accuracy in the range of 72% to 77% where these depended on the time interval.

A malaria warning system was developed in Thailand to improve the previous system in 1984 [82]. The previous system could not provide an accurate report for malaria occurrence because it was formulated based on five years historical data, while malaria occurred in a shorter time period. To solve this problem, a new early detection system that employed *Poisson* model was proposed. The model was developed in three steps, namely model specification, model validation, and model testing. The result showed that the model was suitable to be used for monitoring a weekly malaria incidence at district level.

Chui, *et al* [83] studied the long-term trends of Salmonella hospitalization on U.S. elderly. Hospitalization records of U.S. elderly data from Centers of Medicare and Medicaid were collected in the period of 1991-2004. Data were processed using regression analysis. It was evaluated to observe the long-term trends of Salmonella-related hospitalizations in pre- and post-HACCP (Hazard Analysis and Critical Control Points) periods. The results showed the decreasing prediction rate after 1997 in most division, except South Atlantic, East South Central, and West South Central. Therefore, further study was needed that focused on these three districts. However, because of the different condition in every division, it was important to identify the model that fit to the associated places. Thus, it was able to obtain the better prediction.

Early warning systems for tropical diseases was compared by Chaves and Pascual [84]. Monthly cases of leishmaniasis from January 1991 to December 2001 from Vigilancia de la Salud, Costa Rica were collected for model development. Preprocessing of data was done through normalization using square root method. Linear and non-linear forecasting techniques were applied into the normalized data. The forecast accuracy was measured by using R^2 values. For the linear model, seasonal autoregressive SAR and basic structural model (BSM) were used, while for the non-linear model non-linear forecasting (NLF), generalized additive models (GAM), and feed-forward neural networks (FNN) were used. Based on the result, SAR was determined as the best model.

The application of artificial intelligence (AI) has motivated the use of artificial neural network (ANN) in epidemiological area. Hammad *et al.* [85] worked on the prediction of Schistosoma mansoni infection using ANN. Data from 251 first year students in Egypt were collected in order to predict the infection among the second and third year students. ANN model was built based on backpropagation algorithm. The first year ANN performance was compared with logistic regression, where the result showed that ANN prediction was better than logistic regression. ANN model achieved 83%, while logistic regression achieved 66% only.

An exponential smoothing method was used in forecasting the number of Schistosoma haematobium transmission and intervention impact at Niono, Mali [29]. This method was applied to fit time series data of Schistosoma haematobium based on monthly data in the period of January 1996-June 2004, which it resulted in an online forecasting method. It was encapsulated within a state-space network that was able to show fluctuations of seasonal and inter-annual time series. The result from the forecasting method gave forecast values that provide not only dynamic pattern, but also other supported covariates, such as climate, population and public health intervention. A related research was also conducted in the same district by Daniel *et al.* [86], they constructed a seasonal econometric method by applying multiplicative Holt-Winter's into non-stationery diarrhea, acute respiratory infection, and malaria time series. Their method performed well for diseases with distinct transmission model, and eventually it was used to provide online forecast in that district. The results from both studies [29, 86] supported the improvement of infectious disease management in Mali.

Forecasting technique can also be applied to disease transmission. A mathematical model has been chosen to study Salmonella risk factor in the transmission media (broiler) level at Finlandia [87]. The model consisted of three parts: Primary Production Inference Model (PPIM), Secondary Production Simulation Model (SPSM) and Consumption Inference Model (CIM). WinBUGS and @Risk software were used to develop the system, while Monte Carlo and Markov chain Monte Carlo were chosen as the sampling techniques. The model provided a tool that allows managerial decision to be made based on different scenarios related to economical evaluation.

Various studies in preventing malaria outbreak have been carried out at different locations in the world. An expert system (ES) or advanced DSS have been developed to assist for identification of different species of malaria parasites [33]. Malaria database consists of parasite information such as: identity of parasite, size and shape of red blood cells (RBCs), presence of stained parasite inside RBCs, presence of vacuole inside the parasite, ratio of vacuole to RBCs, presence of pigment granule (Hemozoin) inside the parasite, shape, size and color of Hemozoin pigment granule, and presence number of Merozoites inside the parasite. Two ESs were developed, namely rule based DSS and Bayesian system. The rule base system successfully provides suitable answers based on the input question. The Bayesian system provided a result based on the probability of input estimation. The rule based system was found to provide a more accurate and faster prediction than the Bayesian model.

During the SARS outbreak in 2002-2003, a Bayesian network was applied to find a relationship among the various influenza surveillances [31], which was used for developing a dynamic model. Pediatric data and adult syndromic data in two emergency departments were linked with traditional data i.e. mortality and morbidity. The result showed that the model activated influenza surveillance by predicting the course of influenza epidemic. The model was able to forecast an early influenza outbreak. If the data were incomplete, then a Bayesian method was used as an alternative method for early detection of epidemic outbreak because of its advantage in processing the final result by using any prior data [59].

The similarity of symptoms between dengue and other diseases, especially in the early phase, is often confusing. That has became the motivation for developing a system that could identify dengue fever in the early phase of illness [88]. A total of 1,200 patients were observed for a four week period. Data were processed by using decision tree algorithm. The result showed that, 364 patients were positively dengue RT-PCR, 173 had dengue fever, 171 had hemorrhagic dengue fever, and 20 had dengue shock syndrome. The data were analyzed by a C4.5 decision tree classifier and the results were able to differentiate dengue from non-dengue with 84.7% accuracy.

The aforesaid studies were conducted to predict zoonosis incidence that was based on time series form. Each of the studies focused study on one disease only. Therefore, most of them developed models using a single forecasting method, with the implementation of ARIMA [30, 75-79], regression analysis [28, 32, 80-83], neural network [85], exponential smoothing [29, 86], bayesian [31, 87], and decision tree [88]. In addition, a few researchers stressed the use of multi methods in their model, with Briët, *et al* applied exponential smoothing and ARIMA [35] to predict malaria incidence at Srilanka. Shankar, *et al* emphasized the use of bayesian and rule based [33] to identify various malaria parasites. Whereas, the other two studies used three methods, including Sai, *et al* [34] that apply moving average, exponential smooting, and ARIMA in the prediction schistosomiasis prevelance at China, and work from Chaves and Pascual [84] that compared the used of ARIMA, regression analysis, neural network to analyze tropical diseases. As a conclusion, all the reviewed studies had the same objective in zoonosis prediction even they applied various approaches.

2.4.2 Zoonosis Prediction-Based Spatial-Temporal Data

Disease surveillance is an activity to monitor the spread of diseases for the purpose of controlling the number of incidences and disease outbreak [89]. Spatial-temporal data (data based on geographic area) are widely used in surveillance system. Spatial-temporal model is the extention of spatial model (model base on time series). When data collection for model development is collected based on time and across space, it is called as spatial-temporal model [90].

Spatial data can be represented in GIS (Geographic Information System), in which GIS applications have been used to analyze and evaluate animal diseases and zoonosis outbreak. Pfeiffer *et al.* [15] described the use of GIS application for controlling animal diseases, particularly its possibilities and limitations. Using GIS, disease outbreaks can be mapped into one system. The following examples explore the use of information technique, either temporal or spatial model in disease surveillance, especially related to zoonotic disease.

A computer model was developed to analyze the dynamic spread of fox rabies across the state of Illinois and to evaluate possible disease control strategies [14]. The main concern was focused on the spread of disease from foxes to humans through pet population. The initial variables were determined, such as population densities, fox biology, home ranges, dispersal rates, contact rates, and incubation periods. The resulting system was a GIS computer model that was able to connect data obtained from previous models, fox biology, rabies information and landscape parameters using various hierarchical scales. The system also had a capability to follow the emergent patterns and facilitates experimental stimulus/result data collection techniques. The research concluded that the disease could spread from foxes to humans through the pet population.

In another case, the eradication program of tsetse flies in Zambia, the causal agent of trypanosomiasis, was supported by developing a DSS based on GIS technology [18]. A combination of a tree-based decision-support approaches and the use of Multiple-Criteria Evaluation (MCE), within a GIS, was developed to target the area for tsetse control. This study presented the application of remote sensing incorporated with the environment data to develop a decision support system of tsetse control program.

In the case of Avian Influenza in China, Jinping *et al.* [40] developed a GIS of Avian Influenza transmission model to describe the dissemination of Avian Influenza and to predict future epidemic outbreaks. First, they determined the transmission path of the Avian Influenza epidemic, and then analyzed path. Second, the result was analyzed using stastitical method. Thus, the dissemination of Avian Influenza could be identified and be predicted for the future epidemic situation. The output could assist the Chinese government in making better policy to avoid the Avian Influenza transmission.

The GIS technology also has been applied to assess the risk of wind-borne spread of FMD virus in Australia [44]. The risk of wind-borne infections was ranked and identified by linking an intra-farm virus production model, a wind transport and dispersal model, and an exposure-risk model. Surveillance setting and control management of FMD could be allocated based on the rank of the risk.

Visceral Leishmaniasis diseases generally occur in remote geographical area where health facilities are not established. The diseases that often occur are malaria and other parasitic infections. Due to unavailability of health facilities, it was difficult to make the correct diagnosis. Thus, to overcome this problem, information on geographical distribution in endemic countries was developed. However, because of limited surveillance for obtaining data, GIS was used to find and map the case [45].

The association between environmental and *Schistosoma mansoni* infection in Western Côte D'ivoire was modeled by Beck-Wörner [91]. The model was constructed by using a remotely-sensed digital elevation model (DEM), derived hydrologic features, and Bayesian geostatistica model. Data from 5,448 children in grades 3-5 were collected and grouped by gender and age. Statistical calculation was applied to obtain significant number of infected children. From the data, 1,866 (39.2%) were infected with *S. mansoni*. *S. mansoni* case and indicator of demographic and hydrologic were associated using logistic regression analysis. The result showed good accuracy, and simplicity of data collection was important in broader public-health application. This work emphasized the use of DEM, GIS application, and Bayesian spatial model as an approach to develop the risk profile for other tropical disease that have persisted in the developing countries.

A study in the relationship of global warming with transmission of vectorborne diseases was carried out by Yang *et al.* [92]. Time series data of *Schistosoma Japonicum,* from Jiangsu province in eastern China, for the period 1972-2002 were used as the historical data. Then a modeling analysis was done for estimating the annual growing degree-days (AGDD). The final model included 2 parts, both temporal and spatial. The temporal part had a seasonality component with polynomial order 2 and a periodicity of 3, 6, and 12 months. The spatial part was formed by coordinate of polynomial order 2 added by thin-plate smoothing splines. The forecast for 2003 indicated increasing number of AGDD. This model combined between temporal and spatial components to analyze the effect of temperature to the transmission of schistosomiasis. Based on the results, it was concluded that there was a relationship between temperature changes and level of transmission.

A decision support tool for controlling malaria vector was also conducted in Madagascar [93]. It was based on the epidemic outbreak in the 1980s that killed 30,000 people in the country. The tool was developed to determine areas that required indoor insecticide spraying to kill malaria vector, and hence to prevent the outbreak. Six zones were chosen according to their geographical location in the central highlands and the parasite prevalence determined in 1998. Multicriteria evaluation with weighted linear combination was used as a base to improve action target in determining priority zone. However, the system required further data collection in order to monitor changes and to improve system accuracy.

The high mortality rate of malaria in Kenya had been observed by Li *et al.* [94]. It focused on the transmission of mosquito as disease vector. In those cases, the control strategic could be effective if the distribution of mosquito could be predicted. This study compared the application results between spatial and non-spatial methods. Data was recorded by GPS from 200 houses in the selected area. The mosquitoes in selected houses were surveyed and collected using indoor pyrethrum spray. The samples then were analyzed in the laboratory, where only female mosquitoes were chosen in the study because of their responsibility to the malaria transmission. The results showed that the spatial model was better than non-spatial model. Hence, the result could be used to assist decision making of malaria prevention.

Risk factor identification based on ecological parameter was carried out for prediction of malaria incidence in Koraput district of Orissa, India [95]. The proposed system was constructed using GIS and supported decision making on control strategy. The dataset was derived from remote sensing (IRS-1D/ LISS III), topographic maps (1:50,000), surveys, ground-truth and epidemiological data from the district. The data was later used to construct primary spatial infrastructure. The aim of the GIS was to observe *Anopheles minimum* (malaria vector). The result showed that some parts of Boipariguda and Lamtaput had the largest prevalence of *An. Minimus*. The study was able to determine risk factor based on ecological parameter, where it could be used to develop the decision support for proper control strategy.

INFERNO refers to INtegrated Forecasts and EaRly eNteric Outbreak. Knowledge of infectious disease epidemiology was developed as an adaptive prediction system. The data used for model development were obtained from daily data of cryptosporidiosis in 1993 at Milwaukee, Wisconsin. The system comprises four components: 1) training, 2) warning and flagging, 3) signature forecasting, and 4) evaluation. In the training component, the nature of endemic variation was analyzed, while in component 2 streamline level was quantified to switch the system mode from endemic into epidemic. If the epidemic mode was identified then a forecast of outbreak was made in component 3. In evaluation component, the result in component 3 was analyzed to calculate any uncertainties in the predicted size of the outbreak. The system had been successfully used to predict waterborne outbreak of *cryptosporidiosis* in Milwaukee [96].

Differences between time series and spatial-temporal results tend to be in the type of data and the method used. However, as seen in the reviews, they had the similarity in purposes. Both of groups considered the use of various model in zoonosis prediction. An in-depth study of spatial-temporal model is beyond the scope of this research but is referenced for historical purposes and for showing other approaches beside time series in zoonosis field.

2.5 Decision Support System

The concept of Decision Support System (DSS) is very broad because of the many diverse approaches and wide range of domains in which decisions are made. Power [97] wrote that research on DSS started in 1960s with the development of a model-driven system. It was followed by the development of theory in 1970s, and then the application on financial planning, spreadsheet DSS and Group DSS (GDSS) in the 1980s.

Turban defines DSS as an approach (or methodology) for supporting decision making [98]. DSS [99] also can be defined as a system under the control of one or more decision makers that assists in the activity of decision making by providing an organized set of tools intended to impart structure to portions of the decision-making situation and to improve the ultimate effectiveness of the decision of the outcome. In general, DSS is a computerized system than can assist the user for making decisions.

There are two basic types of DSS, namely model driven and data driven [97].

- Model driven DSS - it is a system that uses a model to perform “what-if” and different analysis in statistical, financial, optimization, or simulation model. In model driven system, there is a need to develop good user interface that can transform the model for easy application by the user. Sensitivity analysis is widely used in models that ask "what-if" question from the assumed condition to determine the impact of changes in one or more parameters.

- Data driven DSS - it is a system that is based on data for assisting in decision making. A large quantity of data is stored in a data warehouse, which the user can access and manipulate the data for their purpose.

The implementation of those two basic types of DSS later give rise to new types of DSS:

- Communication driven DSS - this kind of system is commonly used in group DSS (GDSS). Internet and technology of communication are used to facilitate collaboration and communication among the group member.
- Document driven DSS - within this system, computer storage and technology are used to process input into document retrieval and its analysis. Furthermore, This system is also called text-oriented DSS [100].
- Knowledge driven DSS - in this system, knowledge on particular domain is stored into the database. The knowledge is used by the manager to assist in the process of decision making.

A DSS application can be composed of some subsystems [103]. Despite the variety of DSS applications, basically DSS consists of three primary components [101-102]: database management system, model base management system, and dialog generation and management system. Figure 2.1 presents the DSS components and the interaction between them.

As illustrated in Figure 2.1, there are three basic DSS components which are described as follows:

- Database management system (DBMS)

  Data for DSS design is stored in the DBMS. Data is created, proceeded, and manipulated in this component. DBMS is different from the physical data structure. Hence, users get the information regarding data type and the way for accessing data.

- Model base management system (MBMS)

  Data within DBMS is transformed to create useful information in this system. MBMS is able to assist user in model building. Then, the main function of this component is as a link that separates a model and user application.

- Dialog generation and management system (DGMS)

  The purpose of DGMS is as a user interface in aiding model development. Using this dialog, user can access the DSS application without the need to know the process inside the system. As a user interface, the DGMS is supposed to be user friendly because sometimes the user is a manager who is not a computer expert.

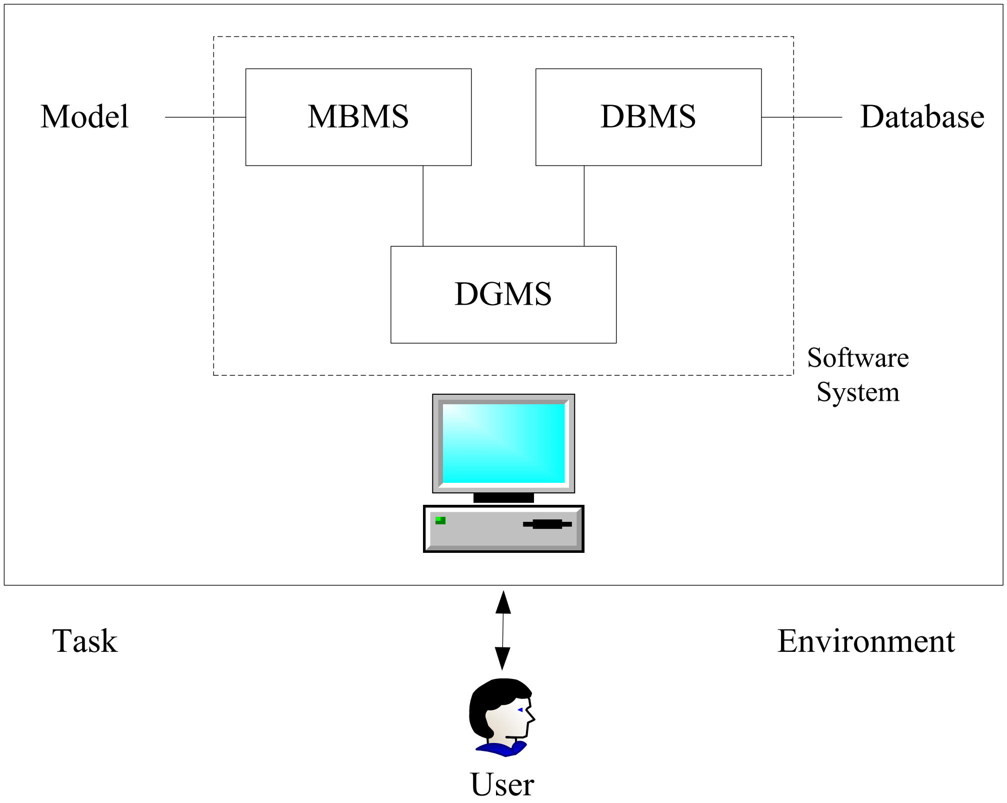


Figure 2.1 DSS Components [101-102]

An advanced DSS has one more component that is known as knowledge-based management subsystem (KBMS) [98]. KBMS provides an expertise component in DSS that is commonly used in expert system or intelligent system. An expert system (ES), also known as a knowledge based system, is a computer program that contains some of the subject-specific knowledge, and the knowledge and analytical skills of one or more human experts [99, 103]. Figure 2.2 shows a schematic view of a DSS.

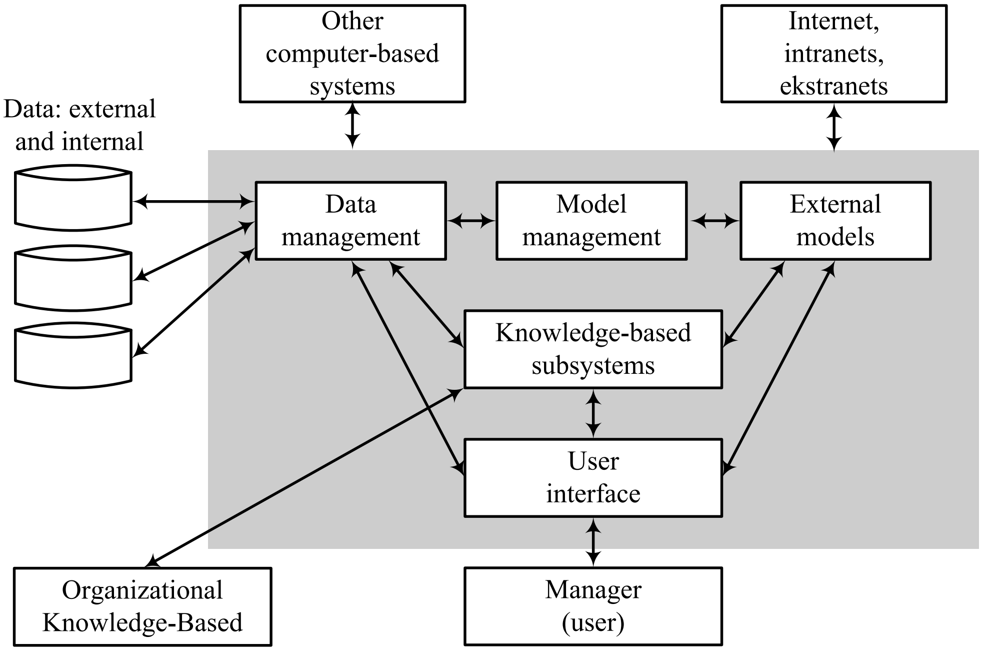

Other computer-based systems
Internet, intranets, ekstranets
Data: external and internal
Data management
Model management
External models
Knowledge-based subsystems
User interface
Organizational Knowledge-Based
Manager (user)

Figure 2.2 A schematic view of DSS (Ref: Turban *et al.*, 2005)

2.6 Identification of the Need for General Framework of Zoonosis Prediction

Section 2.4 summarizes different applications of zoonosis prediction model. The applications were divided based on time series data and spatial-temporal data. This section discusses some of the limitations of the zoonosis prediction model findings.

Temporal-spatial model emphasized research based on is usually developed based on geographical data area constructed using GIS techniques. Unfortunately, GIS has some drawbacks, such as high cost in model development and difficulties to collect spatial data. Most institutions, including health care institutions, store and manage their data in time series form. Thus, purely temporal data are still needed in the surveillance system [59]. Therefore, it is important to choose a suitable technique for temporal surveillance. Different forecasting methods can be applied to support the temporal model. The general aim of those spatial models is to analyze the transmission of diseases and/or the incidence distribution, whereas model based time series are for predicting the incidence of zoonosis.

Table 2.2 provides a various studies on zoonosis forecasting. The models were grouped into two classification, time series and spatial-temporal. A total of 33 studies have been reviewed, out of which 22 are identified as time series (pure temporal) data and 11 as spatial-temporal data. It can be seen that because of the limitation of GIS model, thus most of the related works were based on time series model.

Due to the increasing number of zoonotic diseases, WHO [36] stated that "Efficient early warning and forecasting of zoonotic disease trends through functional surveillance systems is key to effective containment and control". On this basis, there is a need to develop a general zoonosis prediction framework to analyze zoonosis trends and to predict future number of zoonosis incidence. The conclusions of these previous works have brought to light some areas that need to be improved further in zoonosis prediction.

Literature review revealed various examples of different model in zoonosis prediction. These examples provided insights into method used in the model and the number of zoonosis that was covered. According to the literature, most of the existing time series models use a single forecasting method. There are three models that used more than one forecasting method. There is a dependency between dataset and forecasting method where different methods may yield different result [104]. Since several forecasting methods are available [104-107], it makes the selection of method appropriate is challenging. In order to produce more accurate result, it is important to make the right choice. The fittest forecasting method may be chosen when various methods are being applied to the same dataset. Moreover, all related existing models were developed for specific cases where so far none of them could be applied to other dataset with different behavior. Thus, it is necessary to design a model that is flexible enough to be applied in various types of data.

In this research, a DSS framework in zoonosis prediction is proposed. The DSS framework should able to be applied to different zoonosis with various patterns. The framework will apply various forecasting technique on the same set of data, the framework will address different techniques for the same data to obtain the appropriate best fit model. As shown in Table 2.2, the previous studies highlighted the components of data management and model management. In order to design a DSS framework, this research will include DGMS component into the proposed framework. Through DGMS components, user can add the recent data into system. Therefore, the updated data reveal the new prediction results. It can be used to analyze results fluctuation before and after the changes in the database by applying what-if (sensitivity) analysis through the forecast results.

2.7 Chapter Summary

The definition of zoonosis and its impact on human being have been discussed in this chapter. The literature review has provided the basic concept of DSS, including different forecasting technique that will be used in this research to construct a model base management subsystem, which will be included in the DSS framework. A number of studies associated with application of information technology in zoonosis area have been reviewed, and the different information system techniques have been highlighted. Due to the lack of specific DSS studies in the area of zoonosis, the literature review has been conducted to include the application of other information systems in zoonosis.

The main objective focus of reviewing the various related works is to identify any research gaps in order to define this research objective. Hence, reviewed on the literature has been conducted to address the limitations from those previous works and also to present findings of identifies areas for improvement need in zoonosis prediction system. Report on literature reviews were followed by representation of research methodology in Chapter 3. Furthermore, it was used as the research guidance to obtain the final results.

Table 2.2 List of Zoonosis Prediction System

No	Title	Zoonosis	Time Series							Spatial-Temporal
			MA	ES	ARIMA	Reg.	NN	Bayesian	Others	
1	Biosurveillance And Outbreak Detection Using The Arima And Logistic Procedures [75]	SARS			√					
2	Monitoring the SARS Epidemic in China: A Time Series Analysis [30]	SARS			√					
3	Time Series Analysis of Dengue Incidence in Rio de Janeiro, Brazil [76]	Dengue			√					
4	Application of "time series analysis" in the prediction of schistosomiasis prevalence in areas of "breaking dikes or opening sluice for waterstore" in Dongting Lake areas, China [34]	Schistosoma	√	√	√					
5	Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore [77]	SARS			√					
6	Forecasting malaria incidence based on monthly case reports and environmental factors in Karuzi, Burundi, 1997-2003 [78]	Malaria			√					
7	Forecasting malaria incidence from historical morbidity patterns in epidemic-prone areas of Ethiopia: a simple seasonal adjustment method performs best [79]	Malaria			√					

No	Title	Zoonosis	Time Series							Spatial-Temporal
			MA	ES	ARIMA	Reg.	NN	Bayesian	Others	
8	Models for short term malaria prediction in Sri Lanka [35]	Malaria		√	√					
9	The Climatic Factors Influencing The Occurrence of Dengue Hemorrhagic Fever in Thailand [80]	Dengue Hemorrhagic Fever (DHF)				√				
10	Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia II. Weather-based prediction systems perform comparably to early detection systems in identifying times for interventions [81]	Malaria				√				
11	Multivariate Markovian Modeling of Tuberculosis: Forecast for the United States [28]	Tuberculosis				√				
12	Climate Cycles and Forecasts of Cutaneous Leishmaniasis, a Nonstationary Vector-Borne Disease [32]	Leishmaniasis				√				
13	Early Detection of Malaria in an Endemic Area: Model Development [82]	Malaria				√				
14	Geographic variations and temporal trends of Salmonella-associated hospitalization in the U.S. elderly, 1991-2004: A time series analysis of the impact of HACCP regulation [83]	Salmonellosis				√				
15	Comparing Models for Early Warning Systems of Neglected Tropical Diseases [84]	Leishmaniasis			√	√	√			

No	Title	Zoonosis	Time Series							Spatial-Temporal
			MA	ES	ARIMA	Reg.	NN	Bayesian	Others	
16	Comparative evaluation of the use of artificial neural networks for modelling the epidemiology of schistosomiasis mansoni [85]	Schistosoma					√			
17	State–Space Forecasting of Schistosoma haematobium Time-Series in Niono, Mali [29]	Schistosoma		√						
18	Forecasting Non-Stationary Diarrhea, Acute Respiratory Infection, and Malaria Time-Series in Niono, Mali [86]	Malaria		√						
19	The Use of Predictive Models to Manage Risks Caused by Salmonella in Broilers [87]	Salmonellosis						√		
20	Decision support systems to identify different species of malarial parasites [33]	Malaria						√	Rule Based	
21	A Bayesian Dynamic Model for Influenza Surveillance [31]	SARS						√		
22	Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness [88]	Dengue							Decision Tree	
23	A dynamic model of the spatial spread of an infectious disease: the case Of Fox Rabies in Illinois [14]	Fox Rabies								√
24	GIS and multiple-criteria evaluation for the optimisation of tsetse fly eradication programmes [18]	Trypanosomiasis								√
25	Study on transmission model of avian influenza [40]	Avian Influenza								√

No	Title	Zoonosis	Time Series							Spatial-Temporal
			MA	ES	ARIMA	Reg.	NN	Bayesian	Others	
26	An integrated modelling approach to assess the risk of wind-borne spread of foot-and-mouth disease virus from infected premises [44]	Foot and Mouth (FMD)								√
27	Visceral Leishmaniasis: New Health Tools Are Needed [45]	Leishmaniasis								√
28	Bayesian Spatial Risk Prediction of Schistosoma Mansoni Infection In Western Côte D'ivoire Using a Remotely-Sensed Digital Elevation Model [91]	Schistosoma								√
29	A Growing Degree-Days Based Time-Series Analysis for Prediction of Schistosoma Japonicum Transmission In Jiangsu Province, China [92]	Schistosoma								√
30	Determining areas that require indoor insecticide spraying using Multi Criteria Evaluation, a decision-support tool for malaria vector control programmes in the Central Highlands of Madagascar [93]	Malaria								√
31	A study of the distribution and abundance of the adult malaria vector in western Kenya highlands [94]	Malaria								√
32	Geographical information system (GIS) in decision support to control malaria – a case study of Koraput district in Orissa, India [95]	Malaria								√
33	INFERNO: A System for Early Outbreak Detection and Signature Forecasting [96]	Cryptosporidiosis								√

CHAPTER 3
METHODOLOGY

3.0 Chapter Overview

This chapter first discusses the methodology for designing and developing a DSS framework that meets the research objectives. First, an overall research methodology framework is presented in Section 3.1. The research framework methodology consists of four stages. Sections 3.2, 3.3, 3.4, and 3.5 further elaborate each stage individually. Section 3.2 is description of Stage 1 which introduces the background study. Section 3.3 describes Stage 2 for overall model design and explanation of its components, namely database management subsystem, model base management subsystem, and dialog generationn and management subsystem. Stage 3 defines model development based on two case studies that is presented in Section 3.4. The last stage, Stage 4 on model analysis is explored in Section 3.5. Finally, this chapter is closed by summary in Section 3.6.

3.1 Research Framework

The research was conducted with the purpose of designing a general DSS framework to predict seasonal zoonosis incidence on human. To fulfill the research aim, the research framework comprises of four main stages. The research began by studying a problem background from various references. The information obtained was used to design the proposed DSS framework. The framework contains a model for additive time series and a model for multiplicative time series. Once the DSS framework was designed, two selected case studies were applied into the model. Finally, the result of the case studies prediction model were compared

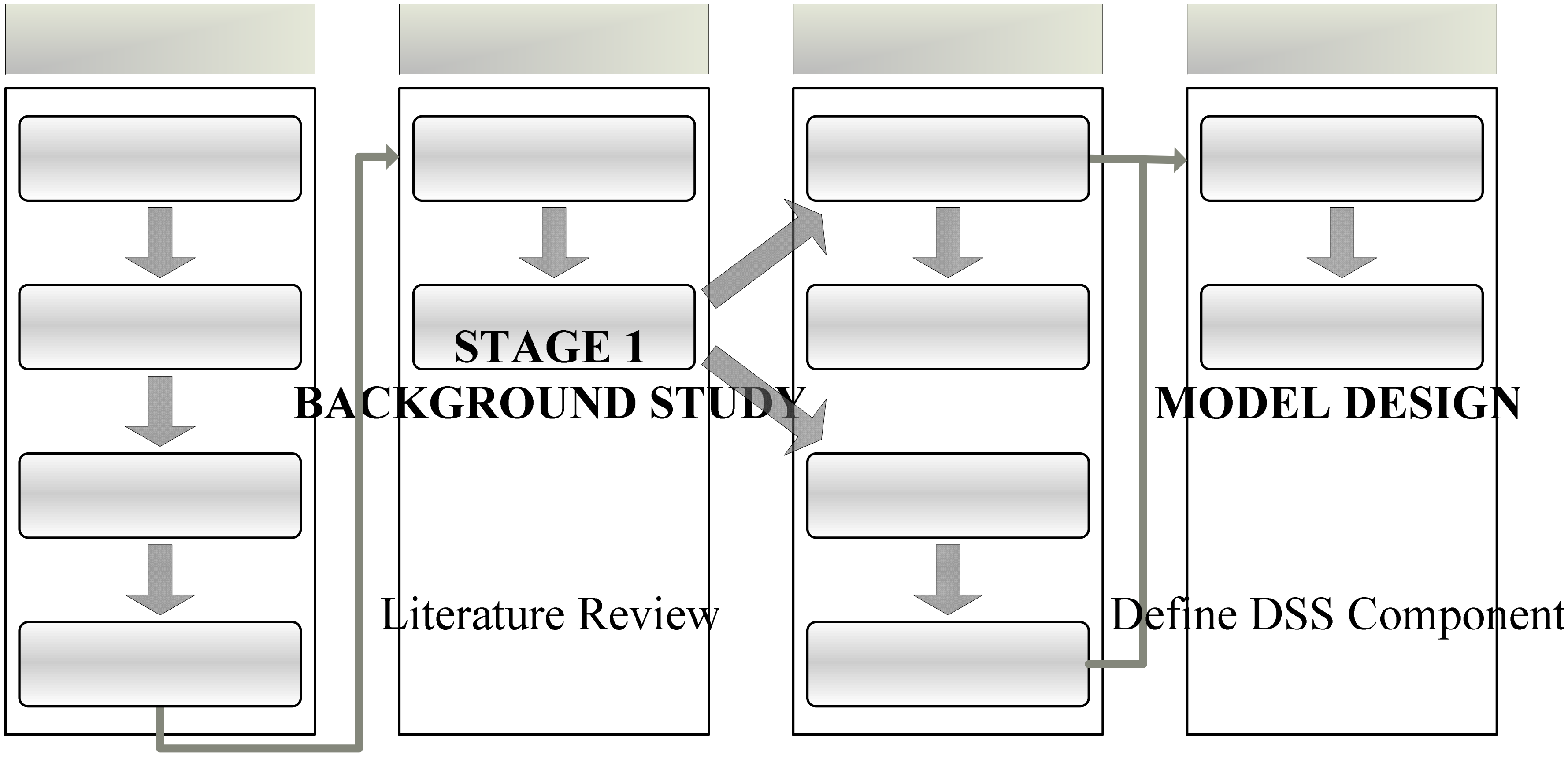


Figure 3.1 Framework of research

The four stages of the research framework are:

- Background study
- Model design
- Model development
- Model analysis

The general research framework is presented in Figure 3.1.

3.2 Stage 1: Background Study

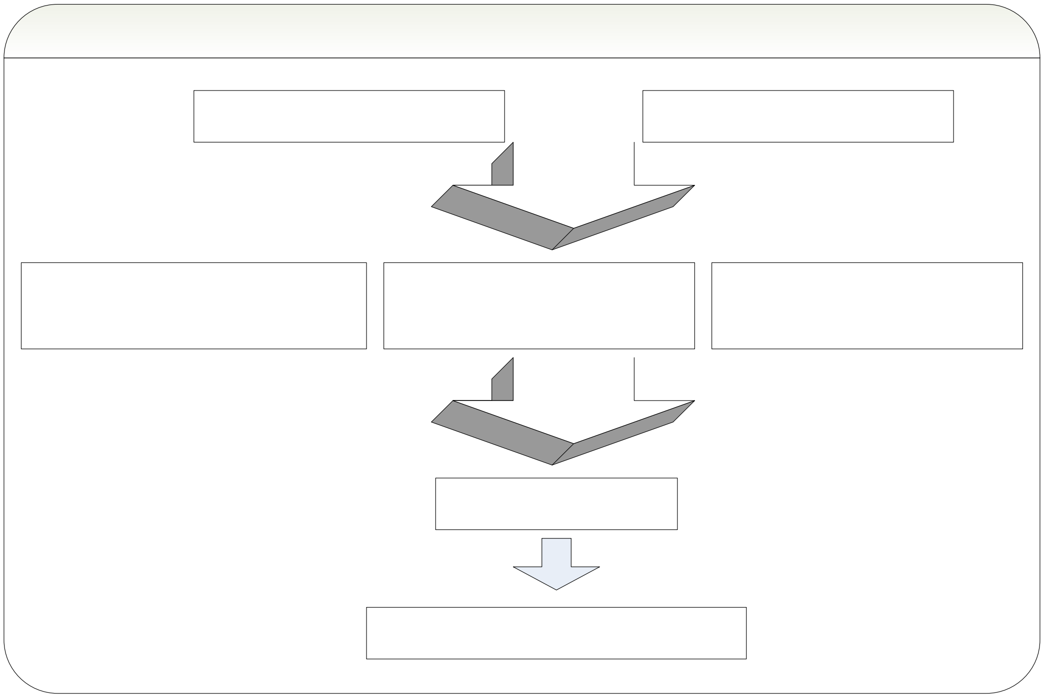


Figure 3.2 Stage 1 – Background Study

Stage 1: Backgrou

The first stage in the framework is research background which consists of three parts, namely problem identification, literature review, and identification of research gap. Chapter 1 and Chapter 2 have provided a discussion on the extended version of stage

Zoonosis

1. Chapter 1 introduced problem identification and research overview. The literature review has provided an overview of zoonosis and its impact on human. Various related work of zoonosis prediction model was reviewed. From the review of related works the research gap that need to be addressed by the proposed framework is then identified. A flow chart outlining the processes in stage 1 is shown in Figure 3.2.

3.3 Stage 2: Model Design

The aim of this research is to develop a DSS framework that is able to predict seasonal zoonosis incidence in human using various methods within the system. The second stage of the framework involves designing DSS model. The DSS model design is based on generic designs that covers two different seasonal components in time series, namely additive seasonal model and multiplicative seasonal model [108-109]. Multiplicative model is a time series model that has a relatively constant trend, while additive model is a time series model that exhibits either upward or downward trend.

Figure 3.3 shows the general DSS framework for model development. The framework is divided into three DSS components. The bulleted number within Figure 3.3 associated to the name of DSS component. Part 1 is the database management subsystem, part 2 is the model management subsystem, and part 3 is the dialog generation and management subsystem. The description of each subsystem and the tools are further discussed in the following subsections.

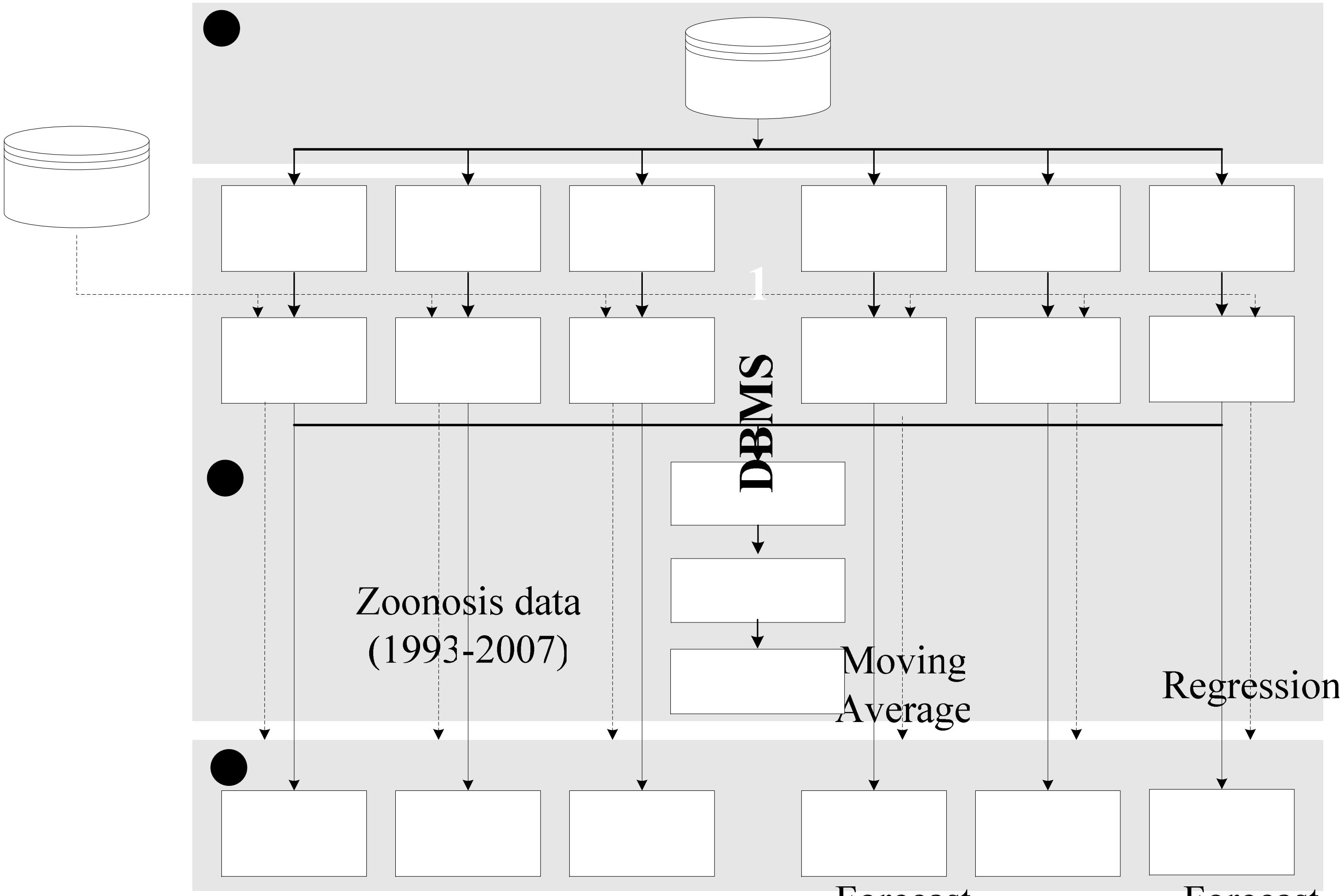


Figure 3.3 The DSS Framework

3.3.1 Database Management Subsystem

This subsystem is as central place to store and to manage DSS database to meet the need of DSS application. The database is managed by database management system (DBMS). DBMS has various function of data management within DSS framework including data definition, data manipulation, data integrity, and data control.

The framework for model development for seasonal zoonosis is time series with either annual additive seasonal time series or multiplicative seasonal time series. Monthly time series incidences for the selected diseases in 1993-2006 were recorded and this yielded 168 monthly data, while the data in 2007 were kept for use in the dialog generation and management susbsytem later. The time series were depicted into chart form in order to analyze the pattern. The result of pattern analysis was used to identify time series seasonality and to determine the appropriate forecasting technique to be used in the model base management subsystem.

In this component, it is important to choose the appropriate case study for model development. The selected case studies are expected to contribute towards achieving the goal of model design. Furthermore, the case studies should be able to identify any special treatment for different zoonosis seasonal time series based on the type of seasonal variation.

3.3.2 Model Base Management Subsystem

The main role of this component is to transform data from database management system into information that can be used in decision making. Several forecasting methods have been developed for prediction purpose.

Forecasting is an action to develop the prediction of future events (forecast) [104]. Using the result, the policy maker can make appropriate decision regarding the future. Forecasting method can be categorized into two general types, namely qualitative method and quantitative method [104]. Qualitative method is used when there is no historical data. It uses experts' opinion to predict future events. On the other hand, quantitative method will be chosen if historical data is available. The quantitative model itself is classified into two groups, namely time series model and causal model.

A time series model is a chronological sequence of observations on specific variables where the variables are based on time function. If relationships between the forecasting variables and the variables are seen as a function other than time, then it is called as causal model. This research is focused on time series model. Some time series statistical approaches have been known, including moving average, exponential smoothing, decomposition, mathematical models, and Box-Jenkins method [104, 110].

In this research, five statistical methods have been chosen within model base management subsystem. The methods were moving average, regression analysis (as a mathematical models), decomposition, Holt-Winter's (as an exponential smoothing), and ARIMA (applied using Box-Jenkins). The selected statistical methods range from the simple technique through the complicated technique. For instance, Neural network is included as a soft computing model [111]. Thus, the results between statistical models and neural network were compared to find the most appropriate method among them.

Before conducting the further process on model base management subsystem, Data were analyzed for trends and seasonality. This analysis helped in choosing an appropriate statistical approach of some statistical method, although it was not necessary for soft computing based model.

The following subsections describe the forecasting techniques used in this research.

3.3.2.1 Moving Average

Moving average is the simplest forecasting method that analyzes the average of the different time series subset [105]. The advantages of this method are that it can smooth the data and present a better illustration of the current trend [105, 112]. Using this method, the future values are obtained by calculating the average of the most recent k data histories. The formula of a simple moving average is:

$$\overline{y_t} = \frac{\left(y_t + y_{t-1} + \ldots + y_{t-n+1}\right)}{n} \tag{3.1}$$

Where y_t is the the value in time period t, and n is the number of terms in each moving average.

Moving average can be obtained by grouping a full time series into a series of number. The average of the first subset is calculated. Then, this subset is moved forward to the new series subset to get the new average. This repeated process is conducted until the end of the data series.

3.3.2.2 Regression Analysis

Regression analysis is a model used to develop the relationship between a dependent variable and one or more independent variables; therefore every component should take the numerical form. Regression model time series is used when the independent variable is time, and the model is focused on predicting future values. This method is able to assemble a linear equation that yields the least square fit of the last m observations [106].

For analyzing time series that exhibits constant seasonal variation, the following general regression formula is used:

$$y_t = TR_t + SN_t + \varepsilon_t \tag{3.2}$$

Where y_t is the observed value in time period t, TR_t is the trend in time period t, SN_t the seasonal factor in time period t, and ε_t is the error term in time period t.

Seasonal factor can be modeled by applying dummy variables. Then SN_t is defined as follows:

$$SN_t = \beta_{s1} x_{s1,t} + \beta_{s2} x_{s2,t} + \ldots + \beta_{s(L-1)} x_{s(L-1),t} \tag{3.3}$$

Where $x_{s1,t}, x_{s2,t}, ..., x_{s(L-1),t}$ are dummy variables, and L is the season used at time series base.

3.3.2.3 Decomposition

Decomposition method is one of the seasonal smoothing methods. The purpose of this method is to distinguish factor of the time series by breaking a series into its components: trend, seasonality, cyclical, and irregular factor [107]. A calculation will be done to obtain the value for each component. These values will be projected forward and they will be reassembled to develop a forecast. The mathematical function for the decomposition approach is:

$$y_t = f(TR_t, SN_t, CL_t, IR_t) \tag{3.4}$$

Where,

y_t = observed value of the time series in time period *t*

TR_t = trend component (or factor) in time period *t*

SN_t = seasonal component (or factor) in time period *t*

CL_t = cyclical component (or factor) in time period *t*

IR_t = irregular component (or factor) in time period *t*

Decomposition method can be divided into two categories of model: multiplicative decomposition and additive decomposition. When a time series exhibits an increasing or a decreasing seasonal variation, then the multiplicative decomposition model will be chosen [104]. Otherwise, the additive decomposition model will be used when the time series display a constant seasonal variation.

3.3.2.4 Holt-Winter's

The Holt-Winter's method is an extension of the simple exponential smoothing that includes seasonality in the approach. In 1957, C.C. Holt suggested simple exponential smoothing. This first model was used to handle non-seasonal time series and had no trend. Next, he formulated a procedure that also covered trends in 1958. In addition, the formula for seasonality was generalized by Winters in 1965, hence the full formula is named as "Holt-Winter's method" [108, 113]. Holt-Winter's method is applicable when the time series has the seasonal component. In addition, this method are simple and need low cost that can be used easily [114].

There are two types of Holt-Winter's method depending on the seasonality variation, namely multiplicative seasonal model and additive seasonal model. Three smoothing constants are determined in this method: level (α), trend (β), seasonality (γ). Every smoothing constant has values between 0 and 1.

- Multiplicative model is used when the data exhibit increasing or decreasing variation. The formulas are as the following:

Level : $$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1}) \tag{3.5}$$

Trend : $$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \tag{3.6}$$

Seasonality : $$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma)S_{t-s} \tag{3.7}$$

Forecast : $$F_{t+m} = (L_t + b_t m)S_{t-s+m} \tag{3.8}$$

- Additive seasonal model is used when time series has a constant variation. The formulas are:

Level : $$L_t = \alpha(Y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1}) \tag{3.9}$$

Trend : $$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \tag{3.10}$$

Seasonality : $S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-s}$ (3.11)

Forecast : $F_{t+m} = L_t + b_t m + S_{t-s+m}$ (3.12)

For equations (3.5) – (3.12), L_t is estimates of the level, b_t is trend of the time series, S_t is seasonal indices, s is the number of periods in one cycle, Y_t is the current values of observations, F_{t+m} is the linear forecast from t onwards, whilst α, β, γ are smoothing parameters of level, trend, and seasonality respectively.

3.3.2.5 ARIMA

This section introduces the basic theory of Autoregressive Integrated Moving Average (ARIMA). This model is selected because of the capability to correct the local trend in data, where the pattern in the previous period can be used to forecast the future. Besides it is a commonly used method that also supports in modeling one perspective as a function of time (in this case, the number of human case) [75]. The general class of ARIMA (*p*,*d*,*q*) is processed as shown in (3.13).

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \quad (3.13)$$

Where *d* is the level of differencing, *p* is the autoregressive order, and *q* is the moving average order [104]. The constant is notated by δ, while ϕ is an autoregressive operator and θ is a moving average operator.

Seasonal ARIMA (SARIMA) is used when the time series exhibits a seasonal variation. A seasonal autoregressive notation (*P*) and a seasonal moving average notation (*Q*) will form the multiplicative process of SARIMA as $(p,d,q)(P,D,Q)_s$. The subscripted letter ‘s’ shows the length of seasonal period. For example, in an hourly data time series *s* = 7, in a quarterly data *s* = 4, and in a monthly data *s* = 12.

In order to formalize the model, the *backshift* operator (*B*) is used. The time series observation backward in time by *k* period is symbolized by B^k, such that $B^k y_t = y_{t-k}$

Formerly, the backshift operator is used to present a general stationarity transformation, where the time series is stationary if the statistical properties (mean and variance) are constant through time. The general stationarity transformation is presented below:

$$z_t = \nabla_s^D \nabla^d y_t = (1-B^s)^D (1-B)^d y_t \tag{3.14}$$

Where *z* is the time series differencing, *d* is the degree of nonseasonal differencing used and *D* is the degree of seasonal differencing used.

Then, the general SARIMA (*p*,*P*,*q*,*Q*) model is

$$\phi_p(B)\phi_P(B^s)z_t = \delta + \theta_q(B)\theta_Q(B^s)a_t \tag{3.15}$$

Where;

- $\phi_p(B) = (1-\phi_1 B - \phi_2 B^2 - ... - \phi_p B^p)$ is the nonseasonal autoregressive operator of order *p*.
- $\phi_p(B^L) = (1-\phi_{1,L}B^L - \phi_{2,L}B^{2L} - ... - \phi_{P,L}B^{PL})$ is the seasonal autoregressive operator of order *P*.
- $\theta_q(B) = (1-\theta_1 B - \theta_2 B^2 - ... - \theta_q B^q)$ is the nonseasonal moving average operator of order *q*.
- $\theta_Q(B^L) = (1-\theta_{1,L}B^L - \theta_{2,L}B^{2L} - ... - \theta_{Q,L}B^{QL})$ is the seasonal moving average operator of order *Q*.
- $\delta = \mu\phi_p(B)\phi_P(B^L)$ is a constant term where μ is the mean of stationary time series.
- ϕ, θ, δ are unknown parameter that can be calculated from the sample data.
- $a_t, a_{t-1}, ...$ are random shocks that are assumed to be independent of each other.

George Box and Gwilym Jenkins studied the simplified step to obtain the comprehensive information of understanding ARIMA model and of using the univariate ARIMA model [104],[107]. The Box-Jenkins (BJ) methodology consists of four iterative steps:

- Step 1: Identification

  This step is focused on the selection of order of regular differencing (*d*), seasonal differencing (*D*), the nonseasonal order of Autoregressive (*p*), the seasonal order of Autoregressive (*P*), the nonseasonal order of Moving Average (*q*) and the nonseasonal order of Autoregressive (*Q*). The number of order can be identified by observing the sample autocorrelations (SAC) and sample partial autocorrelations (SPAC).

In order to get the time series stationary, there are three possible transformations:

1. Regular (nonseasonal) first transformation (1R) is shown in equation (3.16).

$$z_t = y_t - y_{t-1} \tag{3.16}$$

2. First seasonal transformation (1S), is shown in equation (3.17).

$$z_t = y_t - y_{t-L} \tag{3.17}$$

3. First regular and first seasonal transformation (1R1S) is shown in equation (3.18).

$$z_t = y_t - y_{t-1} - y_{t-L} - y_{t-L-1} \tag{3.18}$$

Where;

z_t = Stationary value of CL data at time t.

y_t = Actual CL data at time t.

y_{t-1} = Actual CL data at time t-1.

y_{t-L} = Actual CL data at time t-L.

- Step 2: Estimation

  The historical data are used to estimate the parameters of the tentative model in Step 1.

- Step 3: Diagnostic checking

  Various diagnostic tests are used to check the adequacy of the tentative model.

- Step 4: Forecasting

The final model in step 3 is then used to forecast the time series values.

This approach is widely used to examine the SARIMA model because of its capability to capture the appropriate trend by examining historical pattern. The BJ methodology has two advantages: the ability to extract a great deal of information from the time series using a minimum number of parameters and the capability of handling stationary and nonstationary time series in nonseasonal and seasonal elements [115-116].

The selection of ARIMA model is based on the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) values. These models use Maximum Likelihood principle to choose the highest possible dimension. The determinant of the residual covariance is computed as:

$$|\hat{\Omega}| = \det\left(\frac{1}{T-p}\sum_{t}\hat{\varepsilon}_t \hat{\varepsilon}_t' / T\right) \tag{3.19}$$

Where p is the number of parameters, ε_t is a residual component.

The log likelihood value is computed by a multivariate normal (Gaussian) distribution as:

$$l = -\frac{T}{2}\{k(1+\log 2\pi) + \log|\hat{\Omega}|\} \tag{3.20}$$

Then the AIC and BIC are formulated as [117]:

$$AIC = -2(l/T) + 2(n/T) \tag{3.21}$$

$$BIC = -2(l/T) + n\log(T)/T \tag{3.22}$$

Where l is the value of the log of the likelihood function with the k parameters estimated using T observations and $n = k(d + pk)$. The various information criteria are all based on –2 times the average log likelihood function, adjusted by a penalty function.

3.3.2.6 Artificial Neural Network (ANN)

Soft computing methods has been widely used in various areas [118]. Soft computing is an approach based on the human mind. Therefore, this method is able to adapt to imprecision, uncertainty, partial truth, and approximation [119]. There are several soft computing methods have been known, such as Artificial Neural Network (ANN), fuzzy system, genetic algorithm, and probabilistic reasoning. In this research, neural network is used to learn the historical patterns of zoonosis incidence to forecast future incidence.

Neural networks are techniques to derive a data model inspired from the function of brain and nervous system. Therefore, neural network can be described as a collection of interconnected neurons that work in parallel (learn) to obtain a specific trend in complex data for predicting future values based on the data [120]. This method has been used rather than traditional methods because of the flexibility. With neural network, there is no need to determine a specific assumption for model because neural network model is directly processed from the data [121].

At the beginning, neural network was used to mimic the tasks performed by human brain. Then, it was used to forecast time series data [122]. Typically, neural network consists of an input layer, hidden layers, and an output layer. There are several neural network types: single layer perceptron, linear neuron, multilayer perceptron, competitive networks, self-organizing feature map, and recurrent network.

Neural network processes the calculation that involves training the network with a representative data. The network consists of a number of inputs and outputs. Between these 2 layers, there are hidden layers that consist of some hidden nodes. The number of hidden nodes and layers is empirically determined to optimize the performance of network and obtain a better result.

Multi layer perceptron (MLP) of feed forward neural network is commonly used in some applications [123]. The general architecture of MLP is presented in Figure 2.3.

Based on Figure 3.4, u_n represents nodes ordered in layer, x_n are inputs, o_n are outputs, and $w_{n,n}$ are unidirectional connections with trainable weight. Then, MLP can be formulated as the following:

$$y_t = b_0 + w_{t1}x_1 + w_{t2}x_2 + ... + w_{n1}x_n \tag{3.23}$$

Where y_t is the predicted value at t and w_{tj} is the weight related to the jth input at time t.

The well-known algorithm of ANN training is back propagation network (BPN). In back propagation, the mean square error between calculated output and the desired value is back propagated into the previous layer to minimize error. It is done by adjusting the node weight.

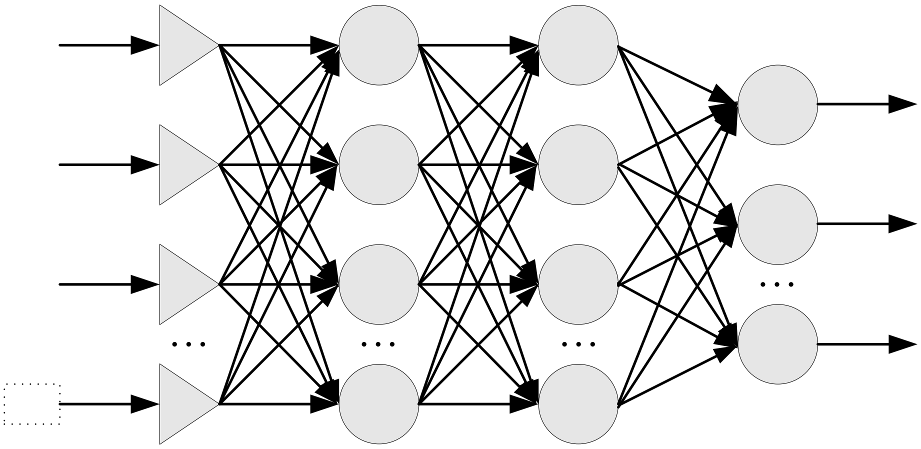


Figure 3.4 Architecture of Multi Layer Perceptron (MLP)

3.3.2.7 ANOVA

In order to determine the fittest method among the forecasting methods, ANOVA were applied (Analysis of Variance) to the forecasting results. The aim of ANOVA is to test the significant difference between group means. ANOVA is commonly used if the user needs to compare performance of more than two parameters. Therefore, the advantage of ANOVA over simple *t*-tests is ANOVA can detect the effect of interaction between variables, and to test more complex hypothesis about the existing

$w_{1,5}$

X_1 u_1 u_5

X_2 u_2 u_6

problem [124]. If the result indicates a significant difference, then it would be followed by post-hoc test to identify which mean of result is different.

ANOVA is a technique that was firstly developed by R.A. Fisher in 1920s. It is used to compare group means [125]. ANOVA uses two hypotheses to determine the result, namely null hypothesis and alternative hypothesis. The hypotheses are presented below:

$H_0 : \mu_1 = \mu_2 = ... = \mu_k$

$H_1 : H_o$ is false where at least one group mean differs.

Where μ_t shows the forecast mean of group *i*.

ANOVA is called as analysis of variance because ANOVA test compares two variance estimations: variance within group and variance between groups.

The formula for variance between groups (s_B^2) is shown below:

$$s_B^2 = \frac{SS_B}{df_B} \tag{3.24}$$

Where s_B^2 represents the sample variance between groups, SS_B is sum of square between groups, and *df* is degree of freedom.

The statistical also called as Mean Square Between (MSB), SS_B can be obtained from equation 3.25 below;

$$SS_B = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \tag{3.25}$$

Where n_i represents the size of group *i*, $\bar{x}_i$ represents the mean of group *i*, and $\bar{x}$ represents a grand mean.

While degrees of freedom can be calculated as,

$$df_B = k - 1 \tag{3.26}$$

Where *k* represents the number of group.

Variance within groups (s_W^2) and the statistical called as Mean Square Within (MSW) is related as follows:

$$s_B^2 = \frac{SS_W}{df_W} \tag{3.27}$$

Where the Sum of Square Within (SS_W) is:

$$SS_W = \sum_{i=1}^{k} (n_i - 1) s_i^2 \tag{3.28}$$

and degrees of freedom is:

$$df_W = N - k \tag{3.29}$$

The values of variance between and variance within yield, *F* statistic is expressed as:

$$F_{stat} = \frac{s_B^2}{s_W^2} \tag{3.30}$$

The F_{stat} is compared with F_{crit} in order to determine whether the null hypothesis is acceptable or not. If the $F_{stat} > F_{crit}$, then the null hypothesis is rejected, otherwise the null hypothesis is accepted.

3.3.2.8 Duncan multiple range test

ANOVA results determine whether there is a significance difference between means of treatment or not. However, there is still important to know which means differ significantly. To determine this, a post hoc test is conducted between pairs of treatment. The use of t-test is not powerful to compare more than two treatments because it can enlarge the overall type I error rate with the number of pairwise comparison [126]. In this research, post hoc test is applied into mean of forecasting result. Duncan multiple range test is chosen because Duncan multiple range test can maintain a low overall type I error [126]. Duncan multiple range test also can be applied in groups application that are not significantly different [127].

When Duncan test uses a studentized range statistic within a multiple stage test, it is called as a multiple range test. A least significant error associated with an increasing number of sample subset means is calculated. The population mean is significantly different if the range of subset is greater than the least significant range. The least significant range is formulated below:

$$R_p = r_p\sqrt{\frac{s^2}{n}} \tag{3.31}$$

Where R_p is the least significant range, r_p is the least significant student range, s^2 is the error mean square from ANOVA result, and n is the number of sample size.

3.3.2.9 Performance of Forecast Results

Coefficient of Variation (CV) is applied to measure the performance of each forecast results. CV is unitless whereas it can be used to compare the various variable in different unit [128]. The suitable methods from the results of Duncan Test was ranked using CV as a ratio of the standard deviation to the mean. The equation of CV is:

$$CV = \frac{\sigma}{\mu} \tag{3.32}$$

Where CV is the coefficient of variation, σ is the standard deviation of the forecast results, and μ is the mean of forecast results.

In this step, the forecasting results of each method were calculated to obtain the value of standard devation σ and μ. The value was used to get the CV value. Once obtaining all CV, it was used to rank among the methods. The methods with the least CV was the best method because it had the less variations than others method. The differences in historical data yields different forecast accuracy whether higher or lower. While in the selection of appropriate method, the user only need to know which method is better than others. Therefore, CV is still a good indicator of comparing forecastibility than error measure that limited by the degree of randomness [129].

3.3.3 Dialog Generation and Management Subsystem

Model management subsystem focuses in model development within DSS, thus the aim of dialog generation and management subsystem is to provide an interaction medium between the user and the model. Through this system, a user can fill in different input and get the result corresponding to the input value. This component links a user to the system that provides a user friendly interface because not all DSS user is an expert. In addition, a user can interact with system to get some recommendation of the related problem [130]. In this research, what-if analysis (sensitivity analysis) was chosen to model dialog generation and management subsystem. What-If analysis (sensitivity analysis) is used to observe how the output varies as influenced by modifying the values in variable input (called as scenario).

As mentioned in Section 2.2, sensitivity analysis is applied to evaluate the system's condition as influenced by parameter changes. In the beginning, a forecasting result was derived based on historical data from 1993-2006. Then, the latest data in 2007 were input into the database. Using the new database, the forecast result was recalculated to get new values. Then, the forecast results due to the recent range (1993-2007) were compared with the forecast results of the previous input (1993-2006). The equation can be written as follows:

$$Sensitivity\,(\%) = \frac{y'(t)_B - y'(t)_A}{y'(t)_B} \times 100 \tag{3.33}$$

Where,

$y'(t)_B$ is forecast number of month t in 2008/2009 based on 1993 – 2007 data

$y'(t)_A$ is forecast number of month t in 2008/2009 based on 1993 – 2006 data

The difference between both results was analyzed to observe the sensitivity of the method. The overall steps of dialog generation and management subsystem especially on what-if analysis results are presented in Chapter 4 and Chapter 5. Once the what-if analysis results are obtained, the DSS flow and interaction between each component were design into a comprehensive Graphical User Interface (GUI) to ease user-system interaction.

A number of tools have been used to design DSS interface. In this research, Visual Basic for Application (VBA) in excel was selected to develop the DSS model that was able to provide an interface between the user and system. VBA for Excel is a programming language that is able to control and automate some function in Excel spreadsheet [131-132]. This tool was selected because of this ability to automate some Excel calculation into a user friendly system. Using VBA for Excel, the user can input the data and process it within MBMS through Excel application. Therefore, VBA for Excel can be applied for DSS based spreadsheet application. The GUI design is described in Chapter 6.

3.3.4 Tools

A number of tools were used to obtain the results. Data of case studies were collected and manipulated using Microsoft Excel 2003. Therefore, the forecasting results were calculated using various softwares. Moving average, regression, and decomposition equations were applied to retrieve the results using Microsoft Excel 2003. For Holt-Winter’s and neural network application, iterations of several models were conducted to achieve the best model using software Statistica 7. For estimating the various ARIMA processes EViews 5.1 econometrics software package was used. The results of each forecasting method were further analyzed using ANOVA, Duncan Multiple Range Test, and What-If Analysis. For this purpose, Microsoft Excel 2003 was comprehensive enough to process these three calculations, where it was only needed to input the formula of each calculation in the same place.

3.4 Stage 3: Model Development

In this stage, case studies were applied into the DSS framework described in Stage 2. Pattern identification of case study is the most important part because the result determines the kind of forecasting approaches to be used in the model base management subsystem. Some forecasting methods differentiate the formula based on time series variation, time series with a seasonal constant pattern and time series with uptrend/downtrend.

In this research, two case studies namely Salmonellosis and Tuberculosis have been selected based on the frequency of incidence and most importantly because of availabilty of data. Salmonellosis and Tuberculosis, specific to the US, have been chosen as the case studies of seasonal zoonosis. This is mainly due to availability of data publicity that could be conveniently accessed through the internet from Centers for Disease Control and Prevention (CDC) websites USA. The case study data were collected from the annual report on summary notifiable diseases that were published by CDC regularly in their website. Besides, the background of case studies selection because of their relative high number of monthly cases compared with other seasonal zoonosis in US.

It was identified that the selected cases have shown high number of monthly incidence as observed from the relatively similar pattern every year. In addition, Salmonellosis exhibits an additive seasonal trend, while Tuberculosis exhibits a multiplicative seasonal trend, thus fulfilling the case requirement. Thus, Salmonellosis was selected as the case study for additive model and Tuberculosis was used for the multiplicative model.

Figure 3.5 shows the processes within stage 3. The selected case studies are analyzed to choose the suitable forecasting method. Referring to Figure 3.5, there is no different formula for regression, moving average, ARIMA, and Artificial Neural Network for the two types of time series data. For these forecasting methods, the same formula and steps can be applied either in additive or multiplicative seasonal model. On the contrary, the trend of time series has an impact in the formulation of the calculation process in decomposition method and Holt-Winter's. There are two types of seasonal time series, namely an additive time series and a multiplicative time series. The additive time series is a time series that shows a relatively constant trend. The multiplicative time series is a time series that exhibits a relatively increasing or decreasing trend. Within model development, the additive decomposition and Holt-

Winter's methods were applied to the additive time series, while the multiplicative decomposition and Holt-Winter's methods were used to the multiplicative time series. Salmonellosis is selected to represent the additive seasonal time series and Tuberculosis is to represent the multiplicative seasonal time series.

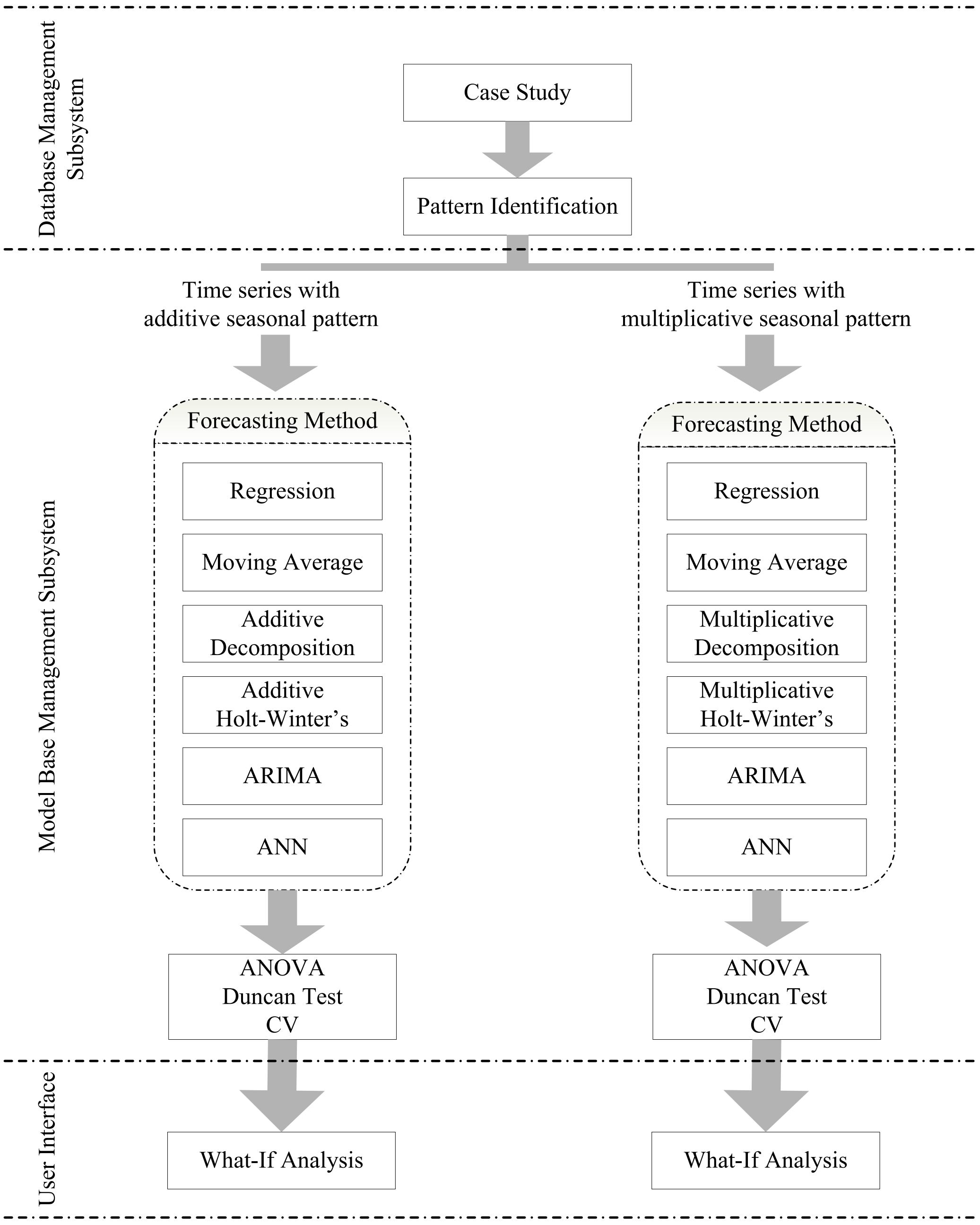


Figure 3.5 Flow of Model Development

After finishing the forecasting process, it was followed by applying ANOVA to check whether there are differences between the forecasting results or not. If the significant different is identified, then Duncan multiple range test is conducted to find the appropriate forecasting method. Once the Duncan result is acquired, Coefficient of Variation (CV) calculation can be used to find the best method. The last step is by applying What-If analysis that is able to measure the fluctuation of forecasting results that is influenced by the changing of data.

The next sections provide a detailed description of Salmonellosis and Tuberculosis, followed by the source from which the time series was collected. These diseases were selected because of the availability of monthly data, their impact to human health, and also the high number of incidence annually.

3.4.1 Salmonellosis

As described by WHO [133],

> "Salmonella is a genus of bacteria that are a major cause of foodborne illness throughout the world. The bacteria are generally transmitted to humans through consumption of contaminated food of animal origin, mainly meat, poultry, eggs and milk."

Salmonellosis or Salmonella was found by Salmon, who was an American scientist [134]. Moreover, Salmonella have been known for over 100 years. Salmonella is a microscopic creature that lives in the intestinal tracks of human and animal. People can be infected with Salmonella if they consume any food that is contaminated with animal faeces. Hence, Salmonella is one of the most common foodborne diseases [135]. When people are infected, they develop diarrhea, fever, and abdominal cramps from 12 to 72 hours after infection. Children are the most likely to get Salmonellosis.

Different Salmonellosis serotypes have been identified. All countries have reported *S. Enteritidis* and *S. Typhimurium* as the most frequently encountered serotypes. This serotype accounts for about 57-67% of total annual reports [136]. Asian epidemiology finds *S. Weltervreden* as the common isolated serotype, this serotype is rare outside Asia [137].

This disease outbreak takes place either in small or in large population and may occur in restaurant, hospital, or some other institutions. Salmonellosis incidences occur worldwide, but the most incidence is reported in North America and Europe [135]. CDC reported that the number of incidence is around 40,000 annually. While in Europe, there is a downward trend from 1995 and this condition differs in each country depending on their control program [138]. In Australia, Salmonellosis incidence exhibits a fairly constant trend since 1998 [139].

Salmonellosis is assumed as a major public health problem in many countries associated to its high cost of impact [68]. It is estimated that the annual total cost related to *Salmonella* around US$ 3 billion in the United States. Unfortunately, there is no published data of the cost of foodborne disease from developing countries.

3.4.2 Tuberculosis

WHO describes Tuberculosis as the following:

> "Tuberculosis, or TB, is an infectious bacterial disease caused by Mycobacterium tuberculosis, which most commonly affects the lungs. It is transmitted from person to person via droplets from the throat and lungs of people with the active respiratory disease."

Tuberculosis is a contagious airborne disease. People who suffer from this disease can spread Tuberculosis germ to others when they cough, sneeze, talk, or spit [140]. The most common Tuberculosis serotype is *Mycobacterium tuberculosis* [141]. The Tuberculosis fact sheet presents that [140, 142]:

- There is one person infected with Tuberculosis every second in the world.
- At around one-third of the world's population is currently infected with Tuberculosis. Every year, 8 million people in the world develop active TB and about 2 million of them are dead.

- It is calculated that 5-10% people who are infected with Tuberculosis (but these people are not infected by HIV) become sick or infectious at some time during their life.

Tuberculosis occurs around the world. There are nearly 9 million new Tuberculosis cases and nearly 2 million TB-related deaths each year [143]. Figure 3.6 shows the global map of Tuberculosis incidence in the world in 2006.

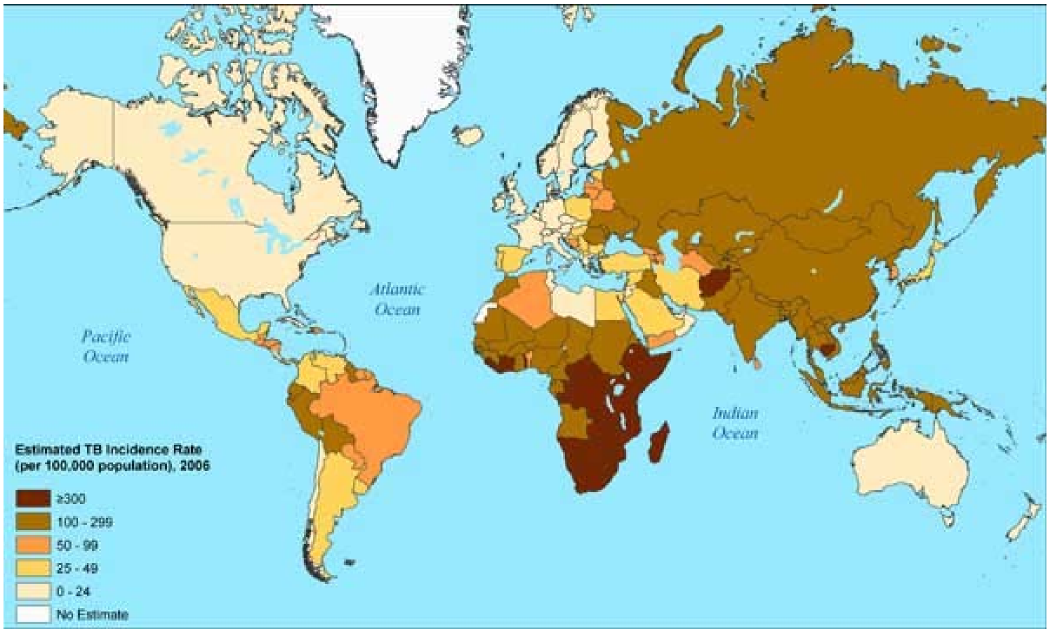


Figure 3.6 Tuberculosis Incidence Rate in 2006 (adapted from WHO report 2008)

3.4.3 Data Collection

Data for the case studies were obtained from the summary of notifiable diseases in United States from the Morbidity and Mortality Weekly Report (MMWR), published by the Center for Disease Control and Prevention (CDC). CDC is one of the operating components of the Department of Health and Human Services, USA that carries the mission of managing different department associated with public health and provides knowledge, information, and tools to the communities to protect their health [144]. The report is published annually and updated regularly to the website.

The Salmonellosis time series exhibits additive seasonal trend where a constant trend has been observed. On the other hand, the Tuberculosis time series displays multiplicative seasonal trend because of its downward trend. The trend of both data is illustrated more in section 4.2 for Salmonellosis and section 5.2 for Tuberculosis. The time series data of both Salmonelosis and Tuberculosis from January 1993 to December 2007 was managed and stored in the database. The data collection and sources for both Salmonellosis and Tuberculosos are presented in Table 3.1. The sample of data is shown in Appendix A.

Table 3.1 List of data collection sources

Data	Sources
Case study in 1993	Summary of Notifiable Diseases, United States 1993 [145]
Case study in 1994	Summary of Notifiable Diseases, United States 1994 [146]
Case study in 1995	Summary of Notifiable Diseases, United States 1995 [147]
Case study in 1996	Summary of Notifiable Diseases, United States 1996 [148]
Case study in 1997	Summary of Notifiable Diseases, United States 1997 [149]
Case study in 1998	Summary of Notifiable Diseases, United States 1998 [150]
Case study in 1999	Summary of Notifiable Diseases, United States 1999 [151]
Case study in 2000	Summary of Notifiable Diseases, United States 2000 [152]
Case study in 2001	Summary of Notifiable Diseases, United States 2001 [153]
Case study in 2002	Summary of Notifiable Diseases, United States 2002 [154]
Case study in 2003	Summary of Notifiable Diseases, United States 2003 [155]
Case study in 2004	Summary of Notifiable Diseases, United States 2004 [156]
Case study in 2005	Summary of Notifiable Diseases, United States 2005 [157]
Case study in 2006	Summary of Notifiable Diseases, United States 2006 [158]
Case study in 2007	Summary of Notifiable Diseases, United States 2007 [159]

3.5 Stage 4: Model Analysis

As illustrated in Figure 3.1, stage 4 is divided into two parts, part 1 is analysis and discussion and part 2 is a report. In stage 3, two diseases were involved as case study approaches. These two cases were discussed with the presentation of each case as a separate and complete study associated to the proposed DSS framework. Once the results were obtained, it was followed with stage 4. A cross case comparison were conducted to the individual case study to determine the similarities and differences among them. The comprehensive discussion can be found in Chapter 6.

3.6 Chapter Summary

This chapter presents the research methodology that comprises into four conceptual stages: 1) Stage 1: Background Study, 2) Stage 2: Framework Design, 3) Stage 3: Model Development; and 4) Stage 4: Model Analysis. Stage 1 introduced a thesis background, including problem identification, literature review, identification of research gaps, and research motivation and contribution. Stage 2 explained the framework design. In this part a DSS framework was presented. The DSS framework has three main DSS components, namely DBMS, MBMS, and DGMS. The framework is designed as a general framework for application in a seasonal time series data. Stage 3 is a result of model development based on DSS framework. Two seasonal zoonosis time series were selected as case studies: Salmonellosis and Tuberculosis. Finally, descriptions of Stage 4 emphasized model analysis from both two case studies.

CHAPTER 4
DSS FRAMEWORK (CASE STUDY ON SALMONELLOSIS)

4.0 Chapter Overview

The research methodology presented in Chapter 3 describes the zoonosis DSS framework in multiplicative seasonal time series and additive seasonal time series. This framework is used iteratively using two case studies, Salmonellosis and Tuberculosis. The result of DSS development for Salmonellosis is reported in this chapter, whilst Chapter 5 reports the result of Tuberculosis. This chapter is organized into 4 sections beginning with introduction to DBMS in Section 4.1, followed by report on MBMS result in Section 4.2, and DGMS result in Section 4.3. For DGMS, the discussion focuses on What-If analysis result. The chapter is concluded by providing a summary in Section 4.4.

4.1 Data Management Subsystem

The dataset for model development was collected from reported cases of Salmonellosis incidence in human, in the United States for a period of 168 months beginning from January 1993 to December 2006. The data were obtained from the summary of notifiable diseases in United States from the Morbidity and Mortality Weekly Report (MMWR) published by the Center for Disease Control and Prevention (CDC). Raw data (as sample in Appendix A) available was processed to form a database, which was subsequently used as input for data management subsystem of the DSS. Using the data, time series analysis was performed. The seasonal variation of the original data is presented as a chart as in Figure 4.1.

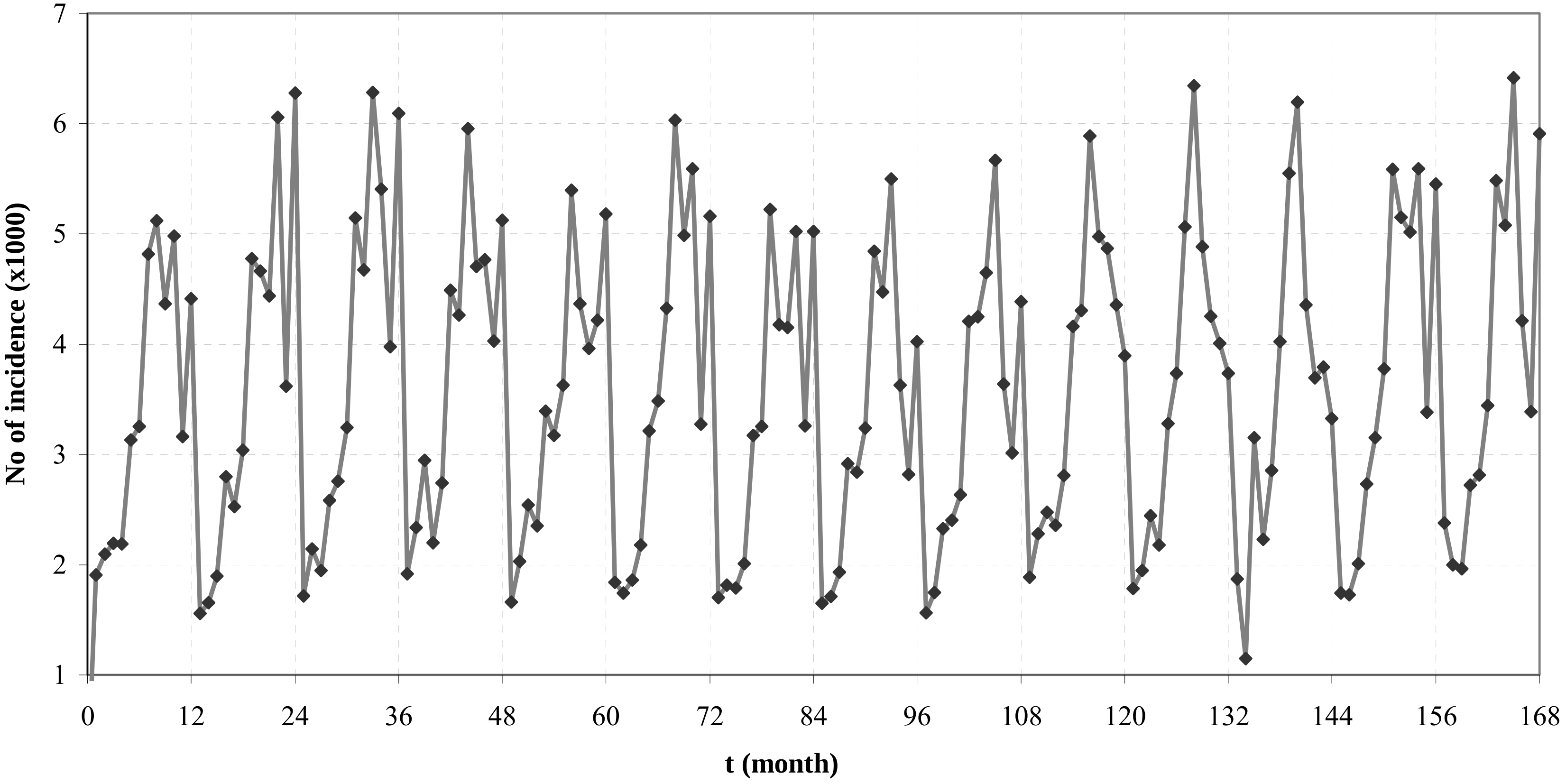


Figure 4.1 Monthly Number of Salmonella Incidence in US (1993-2006)

To present the better illustration of the seasonal component, the time series is grouped based on monthly data and is converted to seasonal stacked chart. Figure 4.2 shows the seasonal stacked line for human Salmonellosis incidence in US from 1993 until 2006. Data in associated month are put together in one group, whereas each group consists of 14 data. In every group, there is a horizontal line that is derived from the data mean in that group. Therefore, the mean lines are able to indicate the level of incidence of each group. The plot shows that the incidence peaked in August while the minimum number of incidence is in January. A time series plot of the historical data shows the seasonal variations, which exhibits similar trend every year.

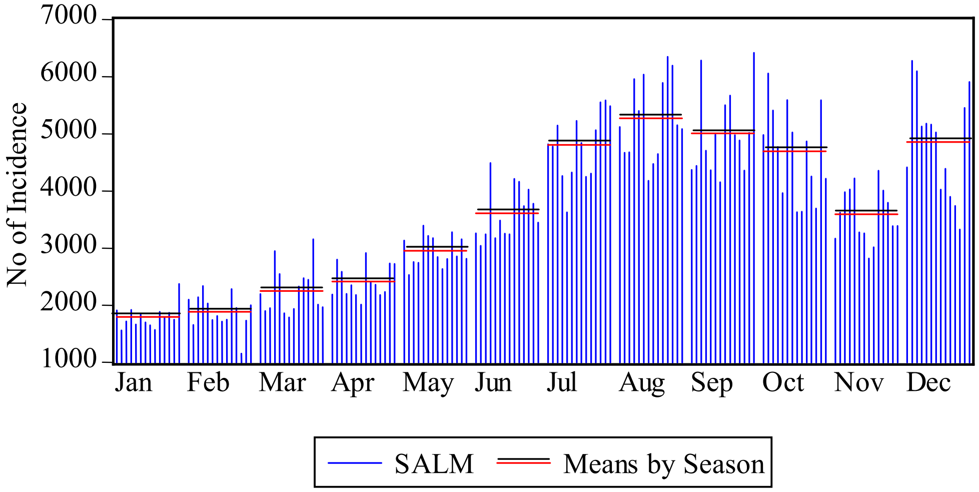


Figure 4.2 Seasonal Stacked Line of US Salmonellosis (1993-2006)

Upon further analysis of time series, it was observed that number of incidents in any month of the year remains significantly constant. Hence, information about time series trend could be used to determine the suitable approach of forecasting monthly diseases incidence in the model management subsystem.

4.2 Model Management Subsystem

In this model management subsystem, fourteen yearly data as presented in data management subsystem are predicted by using five statistical forecasting and soft computing methods. Using the dataset, forecasting models were built using 168 monthly data, and the next 24 months data were forecasted. The results of each method are presented in the following subsections.

4.2.1 Moving Average Result

The first step in the moving average process was to determine the seasonal factor of every month. It was done because of the seasonal variation in the time series. The seasonal basis of the model was based on the annual group. Due to a monthly data, moving average process used twelve historical data where it was started with January as the first month of group and December as the twelfth month of group. Seasonal factors were obtained by dividing the average between annual group mean by the average of the whole data, with the following question. For example, the sequences Salmonellosis data of January 1993 to January 2006 were 1909, 1560, 1716, 1919, 1663, 1840, 1702, 1649, 1566, 1885, 1782, 1870, 1745, 2376. It yielded the January average 1798.714. For the Salmonellosis, the average of the 168 data from January 1993 to December 2006 was 3594.661, thus the seasonal factor of January was 0.500.

Table 4.1 Moving Average Seasonal Factor for Salmonellosis

Month	Seasonal Factor
January	0.500
February	0.524
March	0.626
April	0.673
May	0.821
June	1.004
July	1.336
August	1.467
September	1.393
October	1.305
November	0.999
December	1.351

Table 4.1 lists seasonal factors calculation of moving average using Salmonellosis time series. The values of the seasonal factor represent the level of incidence in every month. According to Table 4.1, the peak of incidence occurs in August with the seasonal factor 1.467 and the lowest incidence occurs in January with the seasonal factor 0.500.

After getting the seasonal factors, the existing dataset need to be deseasonalized first by dividing the actual data with the seasonal factor. In this research, the number of terms in each moving average was 12. In order to calculate moving average, the forecasted results were calculated from the average of the preceding twelve data. This led to the result at the thirteenth step of forecasting model. Finally, the results of the averaging process were multiplied by the seasonal factor to get the final prediction values. The forecast for 2008 and 2009 are presented in Section 4.2.7.

4.2.2 Regression Analysis Result

The time series used in this work were organized into monthly data, so there is a total of 12 seasonal components. Thus, 12 ordinal variables were defined. The twelve seasonal components were determined from the monthly data. Unfortunately, the twelfth month could not provide information like the other 11 months. The twelfth month is used as a baseline for comparison. Thus, a monthly trend variable was applied to analyze the time series.

Table 4.2 presents the use of dummy variables to model the seasonality. In the dummy variables, each variable only have two allowable values, 0 or 1, depend on the month. A month that corresponds to the actual values was assigned as 1 while the rest were assigned as 0. Seasonality was modeled by including dummy variables in the matrix.

Using formula (3.3), the level, trend, and seasonality variables are listed in Table 4.3. As presented in Table 4.3, there are 11 seasonality (monthly) variables based on dummy variables.

The coefficients represented by Table 4.2 are written into equations of formula 4.1 and yield the following:

$$\begin{aligned} y_t &= TR_t + SN_t + \varepsilon_t \\ &= \beta_0 + \beta_1 t + \beta_{s1} x_{s1,t} + \beta_{s2} x_{s2,t} + \beta_{s3} x_{s3,t} + \beta_{s4} x_{s4,t} + \beta_{s5} x_{s5,t} \\ &\quad + \beta_{s6} x_{s6,t} + \beta_{s7} x_{s7,t} + \beta_{s8} x_{s8,t} + \beta_{s9} x_{s9,t} + \beta_{s10} x_{s10,t} + \beta_{s11} x_{s11,t} \\ &= 4815.034 + 0.470t - 3053.470\, x_{s1,t} - 2967.297\, x_{s2,t} - 2603.768\, x_{s3,t} \\ &\quad - 2435.738\, x_{s4,t} - 1902.637\, x_{s5,t} - 1245.321\, x_{s6,t} - 51.220\, x_{s7,t} \\ &\quad + 416.095\, x_{s8,t} + 151.482\, x_{s9,t} - 165.417\, x_{s10,t} - 1264.030\, x_{s11,t} \end{aligned} \tag{4.1}$$

Table 4.2 Design Matrix with Dummy Variables

Time	**J**	**F**	**M**	**A**	**M**	**J**	**J**	**A**	**S**	**O**	**N**
1	1	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0
10	0	0	0	0	0	0	0	0	0	1	0
11	0	0	0	0	0	0	0	0	0	0	1
12	0	0	0	0	0	0	0	0	0	0	0
….											
168	0	0	0	0	0	0	0	0	0	0	0

Applying the use of dummy variables and the regression coefficients yields the forecasted results. For example to calculate the predicted value of January 2003 is described as the following:

$$
\begin{aligned}
y_1 &= TR_t + SN_t + \varepsilon_t \\
&= \beta_0 + \beta_1 t + \beta_{s1} x_{s1,t} + \beta_{s2} x_{s2,t} + \beta_{s3} x_{s3,t} + \beta_{s4} x_{s4,t} + \beta_{s5} x_{s5,t} \\
&\quad + \beta_{s6} x_{s6,t} + \beta_{s7} x_{s7,t} + \beta_{s8} x_{s8,t} + \beta_{s9} x_{s9,t} + \beta_{s10} x_{s10,t} + \beta_{s11} x_{s11,t} \\
&= 4815.034 + (0.470 \times 1) - (3053.470 \times 1) - (2967.297 \times 0) - (2603.768 \times 0) \\
&\quad - (2435.738 \times 0) - (1902.637 \times 0) - (1245.321 \times 0) - (51.220 \times 0) \\
&\quad + (416.095 \times 0) + (151.482 \times 0) - (165.417 \times 0) - (1264.030 \times 0) \\
&= 1762.035
\end{aligned}
$$

Table 4.3 Regression Coefficients for Salmonellosis

Variable	**Coefficient**
Level	4815.034
Trend	0.470
January	-3053.470
February	-2967.297
March	-2603.768
April	-2435.738
May	-1902.637
June	-1245.321
July	-51.220
August	416.095
September	151.482
October	-165.417
November	-1264.030

Thus, the predicted result of January 2003 (t = 1) is 1762.035. The same assumption can be used to obtain the rest of predicted values. It is seen from the calculation, that the seasonal factor of the consecutive month is multiplied by 1, while the entire seasonal factor is multiplied by 0. The results plot of forecasted values up to 2 year ahead (2008 and 2009) can be seen in Appendix B, while the forecasted values for 2008 and 2009 is presented in Section 4.2.7.

4.2.3 Decomposition Result

Based on the current time series (Figure 4.1), which exhibits constant seasonal variation, so additive decomposition is used [104]. The additive decomposition model is

$$y_t = TR_t + SN_t + CL_t + IR_t \tag{4.2}$$

In this case, the component CL_t could be removed from (4.2) because no specific value of CL_t exists. Thus, equation (4.2) can be written as,

$$y_t = TR_t + SN_t + IR_t \tag{4.3}$$

Where y_t is the observed value of the time series in time period t; TR_t is the trend component (or factor) in time period t; SN_t is the seasonal component (or factor) in time period t; CL_t is the cyclical component (or factor) in time period t; and IR_t is the irregular component (or factor) in time period t.

Building a moving average (MA) and a centered moving average (CMA_t) of the time series is the first step in the decomposition method. It was done to eliminate seasonal variation and irregular fluctuation from the data. Then, past values can be smoothed by using this method. The time series consists of 12 monthly data in a year (12 seasons). Then, the centered moving average order 12 is frequently used to estimate the trend cycle in monthly data. All MA must be computed until the last value (y_n) is included in the calculation. Using the MA values, the CMA could be obtained by computing the average of the two adjacent MA values. Equation (4.3) can be rewritten into

$$SN_t + IR_t = y_t - TR_t \tag{4.4}$$

Followed by the estimation of $SN_t + IR_t$ is $sn_t + ir_t$ then

$$sn_t + ir_t = y_t - tr_t = y_t - CMA_t \tag{4.5}$$

The values of $\overline{sn_t}$ were obtained by grouping the values of $sn_t + ir_t$ by months and then calculating the average, $\overline{sn_t}$, for each month. With the L = 12 (number of period a year) then the seasonal factor is:

$$sn_t = \overline{sn_t} - \left(\sum_{t=1}^{L} \overline{sn_t} / L \right) = \overline{sn_t} - 4.719 \tag{4.6}$$

The calculations result of equation (4.6) is given in Table 4.4 and Figure 4.3. Table 4.4 shows the calculation for monthly seasonal factor, and the seasonal factor for each month in the period January - December is presented as a graph in Figure 4.3. It can be observed from Table 4.4 and Figure 4.3 that the peak of incidence occurs in August as the highest seasonal factor among other months (1710.74). Whilst the lowest seasonal factor in January indicates the lowest number of incidence with the seasonal factor -1788.18.

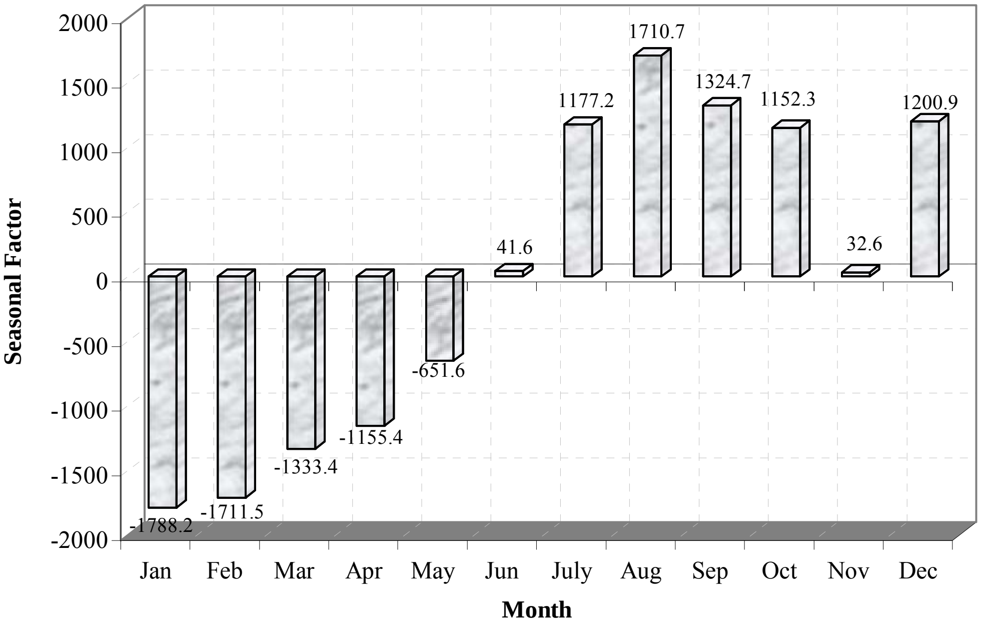


Figure 4.3 Monthly Seasonal factors for US incidence of Salmonellosis

The next step is to calculate the deseasonalized observation (d_t) in time period t as

$$d_t = y_t - sn_t \tag{4.7}$$

Table 4.4 Monthly Seasonal Factor for Salmonellosis Decomposition

Month	***sn_t+ir_t = y_t-(tr_t+cl_t)***													
	Year 1	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Year 6**	**Year 7**	**Year 8**	**Year 9**	**Year 10**	**Year 11**	**Year 12**	**Year 13**	***sn_t***
Jan	-1800.29	-1984.25	-2061.79	-1976.79	-1611.38	-1928.13	-1764.29	-1756.92	-1582.08	-1888.00	-1781.38	-1761.29	-1411.08	-1788.18
Feb	-1683.63	-1574.04	-1660.42	-1559.96	-1763.92	-1776.21	-1695.71	-1557.58	-1239.13	-1770.50	-2514.54	-1734.13	-1781.00	-1711.49
Mar	-1424.63	-1846.33	-1039.04	-1008.54	-1698.29	-1690.13	-1543.17	-992.75	-1067.25	-1289.54	-484.29	-1439.08	-1872.50	-1333.40
Apr	-572.38	-1259.21	-1694.50	-1153.88	-1474.13	-1410.75	-562.17	-921.13	-1208.88	-1527.83	-1360.08	-823.54	-1113.54	-1155.44
May	-906.25	-1074.00	-1125.88	-88.33	-468.79	-222.38	-557.63	-703.83	-864.92	-387.67	-705.83	-462.50	-964.38	-651.62
Jun	-491.83	-596.33	657.38	-314.54	-156.83	-135.83	-100.54	850.67	451.83	91.33	490.38	88.83	-354.42	41.65
July	1363.46	1160.08	1306.71	484.42	126.88	687.58	1841.21	1546.63	863.13	620.63	1419.25	2040.79	1781.88	1177.23
Aug	1696.46	1023.33	819.13	2201.88	1903.46	2401.38	802.58	1179.67	1222.58	2222.75	2732.88	2663.88	1308.38	1710.74
Sep	975.29	771.08	2376.38	977.42	909.88	1352.46	775.63	2189.92	2214.08	1326.88	1274.71	849.42	1166.08	1324.66
Oct	1575.21	2399.04	1476.83	1050.79	542.50	1969.58	1603.79	322.75	182.83	1227.67	612.00	214.29	1741.33	1152.31
Nov	-241.17	-38.50	61.54	278.38	815.08	-337.67	-185.13	-457.33	-445.54	699.63	383.42	279.00	-449.21	32.60
Dec	1041.92	2598.58	2129.29	1405.00	1769.58	1560.67	1593.25	714.96	921.13	239.79	119.08	-190.21	1647.79	1200.94

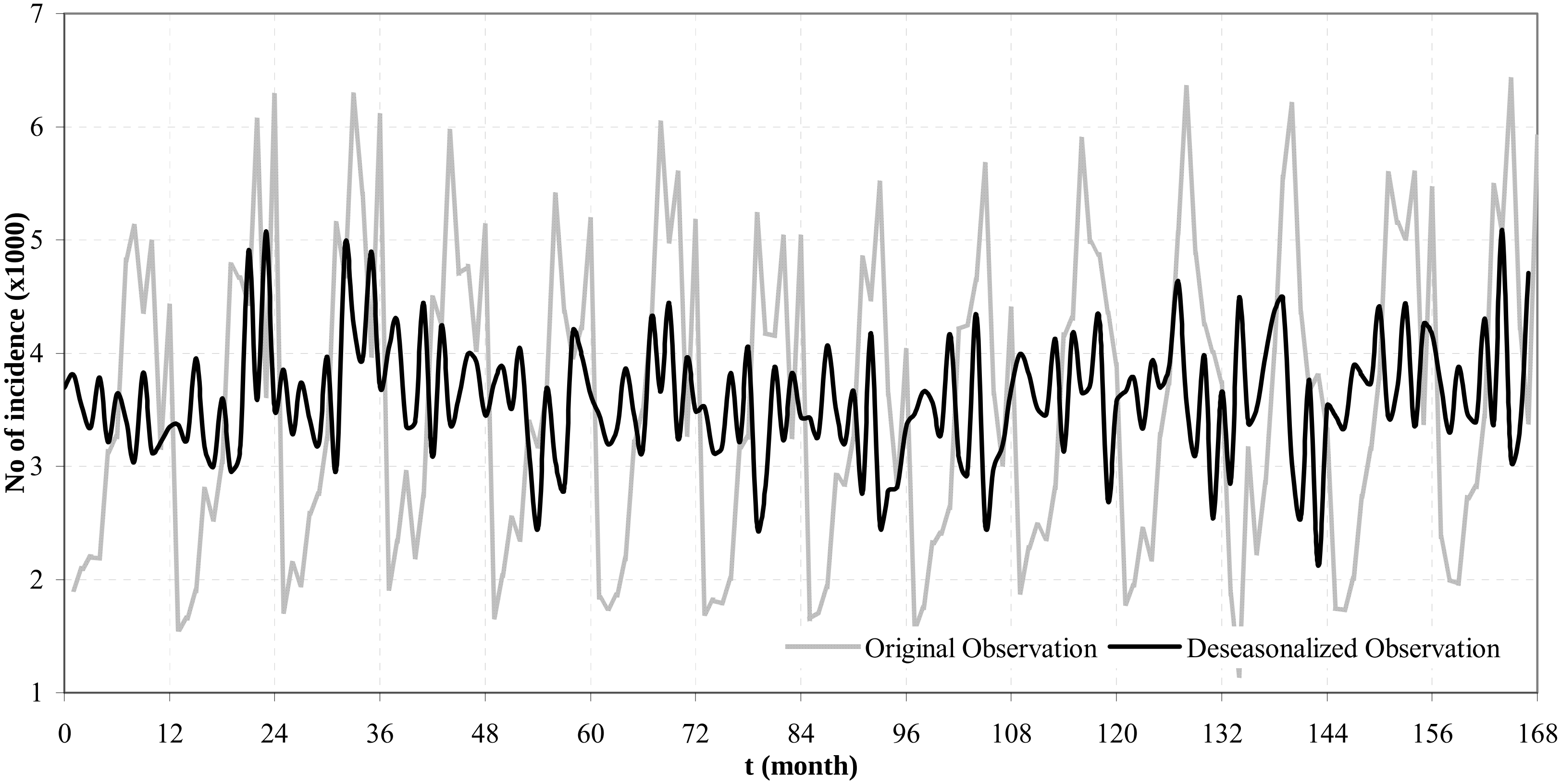


Figure 4.4 Deseasonalized Plot of Salmonellosis

It is necessary to calculate the seasonality first because it is difficult to identify the trend of the time series. Using the seasonal factors, every single point of data can be deseasonalized using equation (4.7). Figure 4.4 presents the deseasonalized chart of the time series. This figure compares the original observation (grey line) and the deseasonalized data (black line). From the figure, the deseasonalized data may remove the seasonal factor for the time series.

Once the deseasonalized data have been calculated, trend of the time series can be determined. The estimation of tr_t for the trend TR_t could be obtained by fitting a regression equation to the deseasonalized data (Figure 4.4). It came out with the following function of tr_t.

$$tr_t = b_0 + b_1 t = 3586.639 - 0.11601t \quad (4.8)$$

By applying equation (4.8) into each point of time, the trend of the related month was yielded. Finally, the final forecasted results could be calculated using equation (4.3). Since the irregular component could not be calculated, it was omitted from the equation (4.3). Therefore, the final results were obtained by using seasonality and trend component in the respective month. The results of the forecasted values in 2008 and 2009 are presented in Section 4.2.7 and the chart of the forecasting results can be found in Appendix B.

4.2.4 Holt-Winter's Result

Based on the time series trend in Figure 4.1, the additive Holt-Winter's method was selected. A Holt-Winter's model required three main smoothing parameters: alpha (α) for level, beta (β) for trend and gamma (γ) for seasonality. Monthly numbers of incidence were used for the forecast model given by equations (3.9) to (3.12). Hence, s was chosen to be 12 (number of months in a year). Method for determining α, β, and γ was based on mean absolute percentage error (MAPE). Different α, β, and γ values were tried and the smoothing parameters that achieved the lowest MAPE was chosen.

To find the suitable smoothing parameters that give the best output, several trial and errors was done by iteration using software Statistica 7, starting parameter 0.01 and increments of 0.01 too. The iteration was stopped at 0.2.

Table 4.5 displays the parameters for ten smallest mean squares from all iterations. Every model presents the value of alpha (α), beta (β), gamma and mean absolute percentage error (MAPE). The best model is achieved by model 2401 which have the lowest MAPE among the model (MAPE = 12.021). The selected parameter was determined by the smallest MAPE with the result: $\alpha = 0.07$, $\beta = 0.01$, and $\gamma = 0.01$.

After selecting the best alpha, beta, and gamma values, the forecast model was built using fourteen years of monthly data. Using the model, a monthly incidence forecast was calculated for 2008 and 2009 as shown in Section 4.2.7 and the actual versus forecasted monthly incidence is found in Appendix B.

Table 4.5 Ten Salmonellosis Model with the Smallest Mean Squares

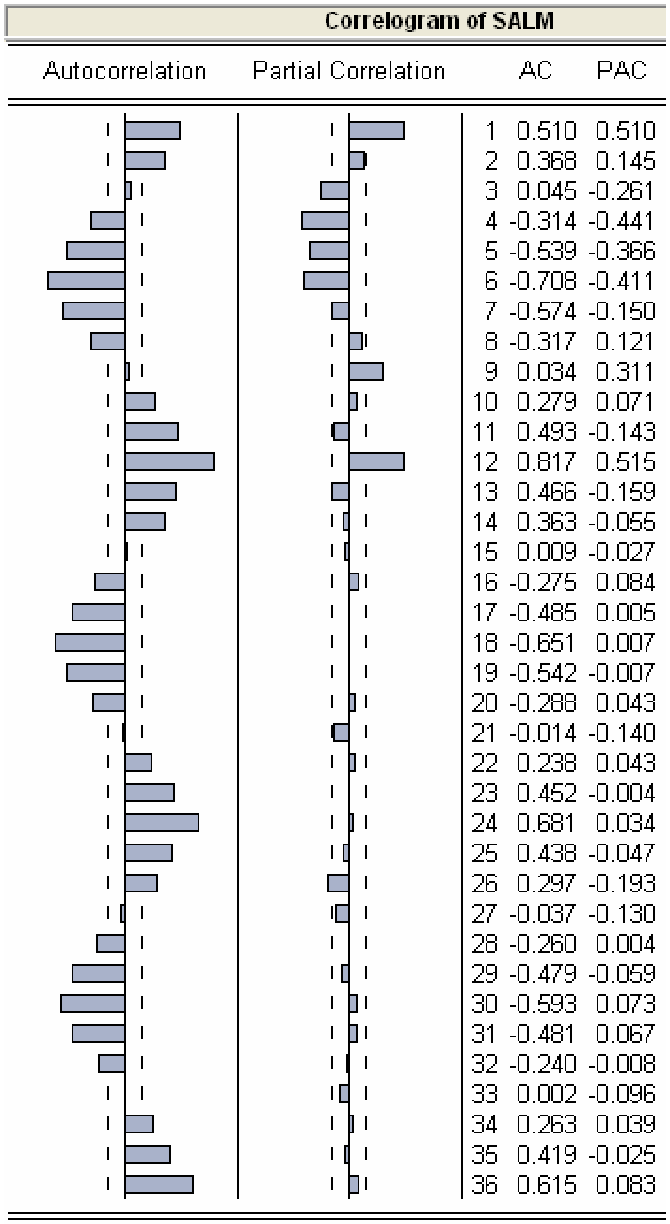
Model	Alpha	Beta	Gamma	Mean Abs % Error (MAPE)
2001	0.06	0.01	0.01	12.025
1601	0.05	0.01	0.01	12.034
2401	**0.07**	**0.01**	**0.01**	**12.021**
1201	0.04	0.01	0.01	12.043
2801	0.08	0.01	0.01	12.033
801	0.03	0.01	0.01	12.069
3201	0.09	0.01	0.01	12.047
2002	0.06	0.01	0.02	12.116
2402	0.07	0.01	0.02	12.111
2021	0.06	0.02	0.01	12.051

4.2.5 ARIMA Result

This section discusses the result of BJ iterative steps to forecast on the available dataset. Software EViews 5.1 was used for ARIMA model development.

4.2.5.1 Identification

According to BJ methodology introduced in Section 2.4.5, the first step in model development is to identify the dataset. The ARIMA is modeled based on the time series stationary, it means that the time series mean, variance, and covariance remain constant over time. In this step, sample autocorrelations (SAC) and sample partial autocorrelations (SPAC) of the historical data were plotted to observe the pattern.



Correlogram of SALM

	AC	PAC
1	0.510	0.510
2	0.368	0.145
3	0.045	-0.261
4	-0.314	-0.441
5	-0.539	-0.366
6	-0.708	-0.411
7	-0.574	-0.150
8	-0.317	0.121
9	0.034	0.311
10	0.279	0.071
11	0.493	-0.143
12	0.817	0.515
13	0.466	-0.159
14	0.363	-0.055
15	0.009	-0.027
16	-0.275	0.084
17	-0.485	0.005
18	-0.651	0.007
19	-0.542	-0.007
20	-0.288	0.043
21	-0.014	-0.140
22	0.238	0.043
23	0.452	-0.004
24	0.681	0.034
25	0.438	-0.047
26	0.297	-0.193
27	-0.037	-0.130
28	-0.260	0.004
29	-0.479	-0.059
30	-0.593	0.073
31	-0.481	0.067
32	-0.240	-0.008
33	0.002	-0.096
34	0.263	0.039
35	0.419	-0.025
36	0.615	0.083

Figure 4.5 SAC and SPAC Correlogram of Salmonellosis Original Data

Figure 4.5 shows the correlogram of original time series. Since the pattern could also be observed from the results, only three periodical data were plotted to confirm the pattern, hence there is no need to display all periodic. The result is shown as time series correlogram in Figure 4.5. The correlogram is divided into five parts. The first part presents the level of Autocorrelation (AC) and Partial Correlation (PAC). The dash line indicates the critical boundaries of correlation values. It is followed by the numbering list (1, 2, …, 36) that shows list of 36 lags. The last two parts are the values of AC and PAC in each lag. Both critical boundaries and AC/PAC values present an availability of correlation in a specific lag, where if the value of AC and PAC lag is greater than |0.2| then there is correlation in that lag. The time series is stationary if no significant correlation in the lags.

Based on Figure 4.5, it is shown that the time series correlogram likely indicates non-stationarity since the series does not appear vertical along the Y-axis, especially in the SAC correlogram. Besides, the non-stationarity of time series can be identified from the values of SAC and SPAC, if the values > |0.2| then the time series is likely not stationarity [104]. According to Figure 4.5, values of SAC at most lag are more than |0.2|. In the SPAC, the first periodic exhibits non-stationarity with the SPAC values are greater |0.2| at lag 1, 3, 4, 5, 6, 9, 12, whilst the next period of SPAC are stationary because the values are smaller than |0.2|. The need of time series differencing also can be determined from Augmented Dickey-Fuller (ADF) unit root test.

Therefore, the ADF unit root test was performed to determine whether a data differencing was needed or not [117]. The null hypothesis of the ADF t-test is:

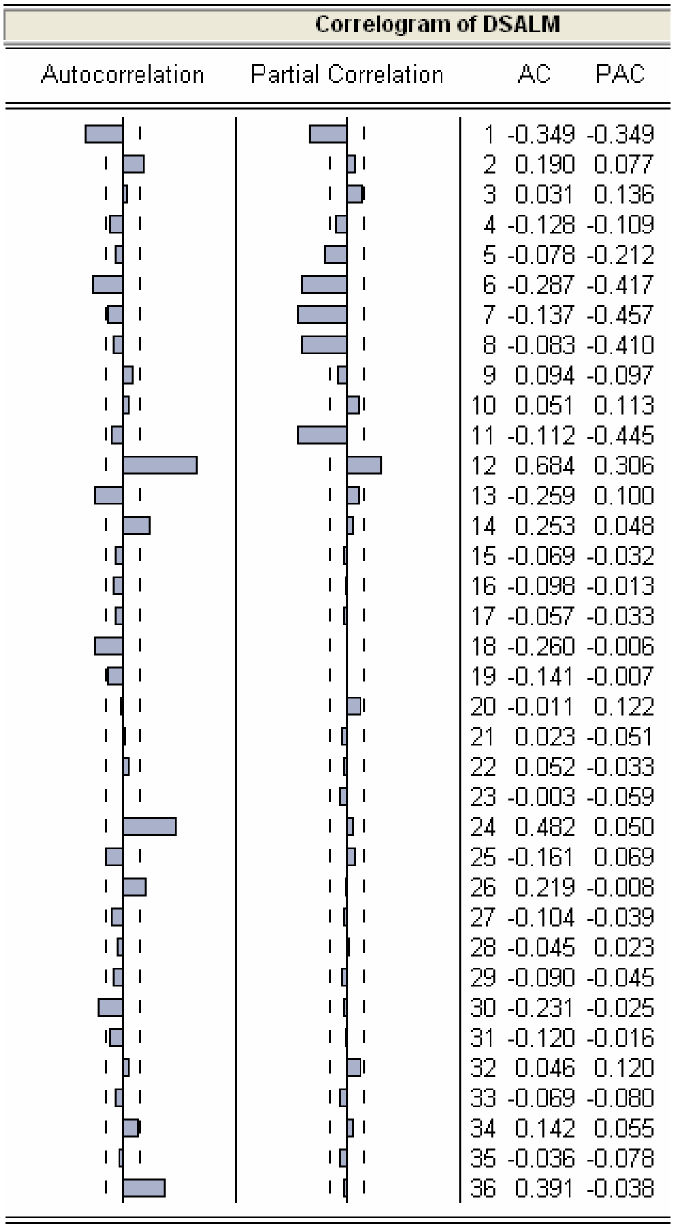
- H_0 : θ = 0, the data need to be differenced to make it stationary, versus the alternative hypothesis of
- H_1 : θ < 0, the data are stationary and do need to be differenced.

The result was compared with 1%, 5%, and 10% critical values to indicate non-rejection of the null hypothesis. The ADF test statistic value has a t-Statistic value of -1.779 and the one-sided p-value is 0.389. The critical values reported at 1%, 5%, and 10% are -3.476, -2.882, -2.578. This shows that t_α value is greater than the critical values, which provides evidence not to reject the null hypothesis of a unit root; then the time series need to be differenced. Therefore, the regular differencing and seasonal differencing was applied to the original time series.

In order to make the time series stationary, regular differencing and seasonal differencing were applied to the original time series. The ADF test was also applied for both of them. The result shows that the critical values of the regular differencing is -14.171 and for the seasonal differencing is -12.514. The one-sided *p*-value for both differencing is 0.000. It provided evidence to reject the null hypotheses and indicated the stationarity of the time series.

Figures 4.6 and Figure 4.7 shows the correlogram of Salmonellosis regular time series differencing and seasonal time series differencing. In these figures, autocorrelation and partial correlation plot for each time lag is presented next to each unit in column SAC and SPAC. Values of SAC and SPAC those are greater than |0.2| indicates a spike. The selection of whether to use regular or seasonal differencing was based on the correlogram.

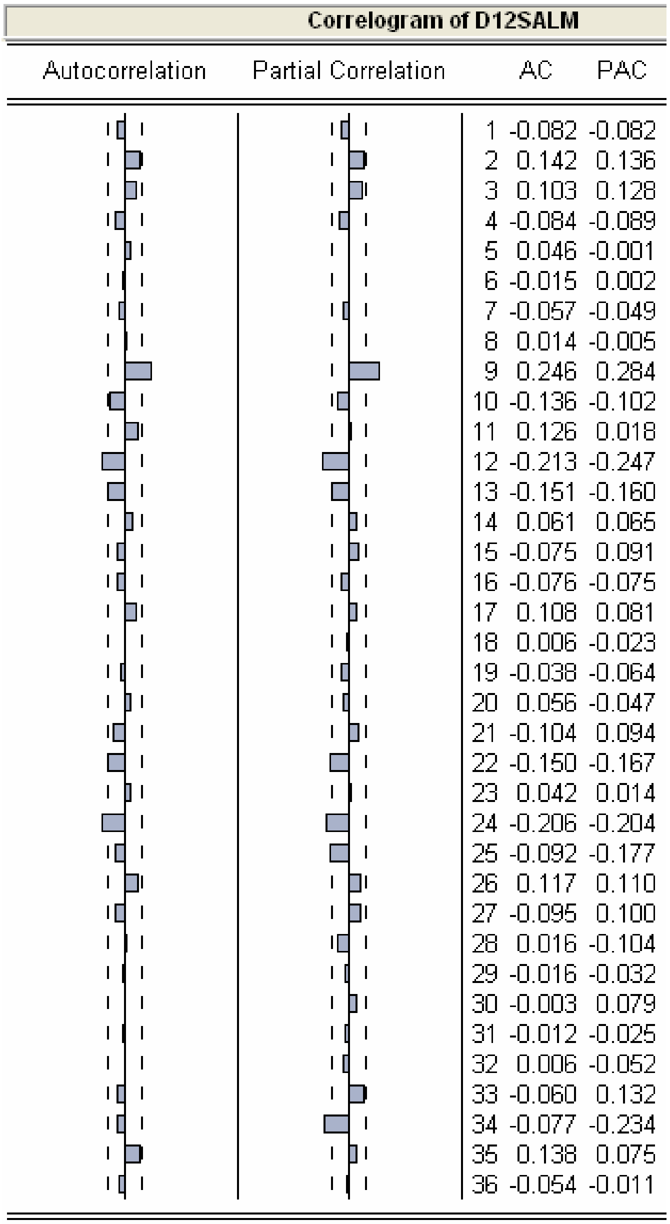
Based on the regular correlograms shown in Figure 4.6, the SAC values at all seasonal lags (lag 12, 24, and 36) > |0.2| and some SPAC values in the first period are greater than |0.2| (lag 1, 5, 6, 7, 8, 11, 12). For the seasonal differencing (Figure 4.7), both SAC and SPAC tend to be stationary with only a few spike indicated there rather than in regular differencing with SAC spikes are at lag 9 and lag 24, and SPAC spikes are at lag are at lag 9, lag 12, and 24. As such, a seasonal differencing of Salmonellosis data is chosen instead of regular differencing for the model development.



Correlogram of DSALM

	AC	PAC
1	-0.349	-0.349
2	0.190	0.077
3	0.031	0.136
4	-0.128	-0.109
5	-0.078	-0.212
6	-0.287	-0.417
7	-0.137	-0.457
8	-0.083	-0.410
9	0.094	-0.097
10	0.051	0.113
11	-0.112	-0.445
12	0.684	0.306
13	-0.259	0.100
14	0.253	0.048
15	-0.069	-0.032
16	-0.098	-0.013
17	-0.057	-0.033
18	-0.260	-0.006
19	-0.141	-0.007
20	-0.011	0.122
21	0.023	-0.051
22	0.052	-0.033
23	-0.003	-0.059
24	0.482	0.050
25	-0.161	0.069
26	0.219	-0.008
27	-0.104	-0.039
28	-0.045	0.023
29	-0.090	-0.045
30	-0.231	-0.025
31	-0.120	-0.016
32	0.046	0.120
33	-0.069	-0.080
34	0.142	0.055
35	-0.036	-0.078
36	0.391	-0.038

Figure 4.6 SAC and SPAC Correlogram of Salmonellosis Regular Differencing



Correlogram of D12SALM

	AC	PAC
1	-0.082	-0.082
2	0.142	0.136
3	0.103	0.128
4	-0.084	-0.089
5	0.046	-0.001
6	-0.015	0.002
7	-0.057	-0.049
8	0.014	-0.005
9	0.246	0.284
10	-0.136	-0.102
11	0.126	0.018
12	-0.213	-0.247
13	-0.151	-0.160
14	0.061	0.065
15	-0.075	0.091
16	-0.076	-0.075
17	0.108	0.081
18	0.006	-0.023
19	-0.038	-0.064
20	0.056	-0.047
21	-0.104	0.094
22	-0.150	-0.167
23	0.042	0.014
24	-0.206	-0.204
25	-0.092	-0.177
26	0.117	0.110
27	-0.095	0.100
28	0.016	-0.104
29	-0.016	-0.032
30	-0.003	0.079
31	-0.012	-0.025
32	0.006	-0.052
33	-0.060	0.132
34	-0.077	-0.234
35	0.138	0.075
36	-0.054	-0.011

Figure 4.7 SAC and SPAC Correlogram of Salmonellosis Seasonal Differencing

4.2.5.2 Parameter Estimation

Different ARIMA models were applied to find the best fitted model. The most appropriate model was selected by using Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) values. The best model was determined from the minimum BIC and AIC. Table 4.6 presents the results of estimating the various

ARIMA processes for the seasonal differencing of Salmonellosis human incidence using the EViews 5.1 econometric software package.

Table 4.6 Estimation of Selected SARIMA Model

No	Model Variable	BIC	AIC	Adj. R^2
1	C, AR(9), SAR(12), SAR(22), SAR(24), MA(9), SMA(12), SMA(24)	15.614	15.431	0.345
2	AR(9), SAR(12), SAR(22), SAR(24), MA(9), SMA(12), SMA(24)	15.587	15.427	0.342
3	AR(9), SAR(12), SAR(24), MA(9),SMA(12), SMA(22), SMA(24)	15.583	15.423	0.345
4	AR(3), AR(9), SAR(12), MA(3), SMA(24)	15.394	15.286	0.449
5	AR(3), AR(9), SAR(12), MA(14), SMA(24)	15.351	15.243	0.472
6	AR(3), AR(9), SAR(12), MA(24)	15.368	15.282	0.447
7	AR(9), SAR(12), MA(3), SMA(24)	15.359	15.273	0.452
8	AR(9), SAR(12), MA(14), SMA(24)	15.331	15.245	0.468
9	AR(9), SAR(12), MA(3), MA(14), SMA(24)	15.352	15.245	0.471

The AIC and BIC are commonly used in model selection, whereby the smaller value is preferred [160]. From Table 4.6, model 8 is selected as the fittest model among them. Model 8 has the smallest value of BIC and a relatively small AIC. It also has large adjusted R^2.

The model is AR(9), SAR(12), MA(14), SMA(24) that also can be notated as $ARIMA(9,0,14)(12,1,24)_{12}$. The notation consists of two parts, the first part is a regular component and the second notation is a seasonal component. The subscript number presents the number of lags in one period, whereas '12' shows that there is twelve lags in one period. Each component is divided into three parts, the first part is autoregressive (AR) component, the second is the differencing level, and the last part is the moving average (MA) component. $ARIMA(9,0,14)(12,1,24)_{12}$ shows AR(9), MA(14) in the regular components and SAR(12), SMA(24) in the seasonal components, where number between bracket presents lags. Because of the seasonal differencing used in model, a differencing level of the regular component is 0 and on the contrary a differencing level of the seasonal component is 1.

To produce the model, the separated non-seasonal and seasonal models were computed first. It was followed by combining these models to describe the final model.

- Step 1: Model for non-seasonal level

  AR (9) : $z_t = \delta + \phi_9 z_{t-9} + a_t$ (4.9)

  MA(14) : $z_t = \delta + a_t - \theta_{14} a_{t-14}$ (4.10)

- Step 2: Model for seasonal level

  AR (12) : $z_t = \delta + \phi_{1,12} z_{t-12} + a_t$ (4.11)

  MA (24) : $z_t = \delta + a_t - \theta_{2,12} a_{t-24}$ (4.12)

- Step 3: Combining (4.9) – (4.12) to arrive at (4.13).

$$z_t = \delta + \phi_9 z_{t-9} - \theta_{14} a_{t-14} + \phi_{1,12} z - \theta_{2,12} a_{t-24} + a_t \qquad (4.13)$$

Hence, δ value is 0.44, less than |2| and statistically not different from zero, then δ was excluded from the model. When an autoregressive and a moving average model are presented in the non-seasonal or seasonal level, the followings are used in the multiplicative terms [104]:

- $\phi_9 z_{t-9}$ and $\phi_{1,12} z_{t-12}$ were used to form the multiplicative term $\phi_9 \phi_{1,12} z_{t-21}$
- $-\theta_{14} a_{t-14}$ and $-\theta_{2,12} a_{t-24}$ were used to form the multiplicative term $\theta_{14} \theta_{2,12} a_{t-38}$

The model was derived using the multiplicative form as described in (4.14):

$$\begin{aligned} z_t &= \phi_9 z_{t-9} - \theta_{14} a_{t-14} + \phi_{1,12} z_{t-12} - \theta_{2,12} a_{t-24} + \phi_9 \phi_{1,12} z_{t-21} + \theta_{14} \theta_{2,12} a_{t-38} + a_t \\ z_t - \phi_9 z_{t-9} - \phi_{1,12} z_{t-12} - \phi_9 \phi_{1,12} z_{t-21} &= \theta_{14} a_{t-14} + \theta_{2,12} a_{t-24} - \theta_{14} \theta_{2,12} a_{t-38} + a_t \end{aligned} \qquad (4.14)$$

The backshift operator (B) was applied in (4.14) yield:

$$\begin{aligned} z_t - \phi_9 B^9 z_t - \phi_{1,12} B^{12} z_t - \phi_9 \phi_{1,12} B^{21} z_t &= \theta_{14} B^{14} a_t + \theta_{2,12} B^{24} a_t - \theta_{14} \theta_{2,12} B^{38} a_t + a_t \\ (1 - \phi_9 B^9 - \phi_{1,12} B^{12} - \phi_9 \phi_{1,12} B^{21}) z_t &= (1 + \theta_{14} B^{14} + \theta_{2,12} B^{24} - \theta_{14} \theta_{2,12} B^{38}) a_t \end{aligned} \qquad (4.15)$$

From the computation, the parameter results are AR(9) = 0.154, SAR(12) = -0.513, MA(14) = 0.255, SMA(24) = -0.860. The estimated parameters were then substituted into (4.15) to form the final model (4.16), which is expressed as follows:

$$\begin{aligned}&(1-0.154B^{9}+0.513B^{12}+0.078B^{21})z_t \\ &=(1+0.255B^{14}-0.860B^{24}-0.219B^{38})a_t\end{aligned} \tag{4.16}$$

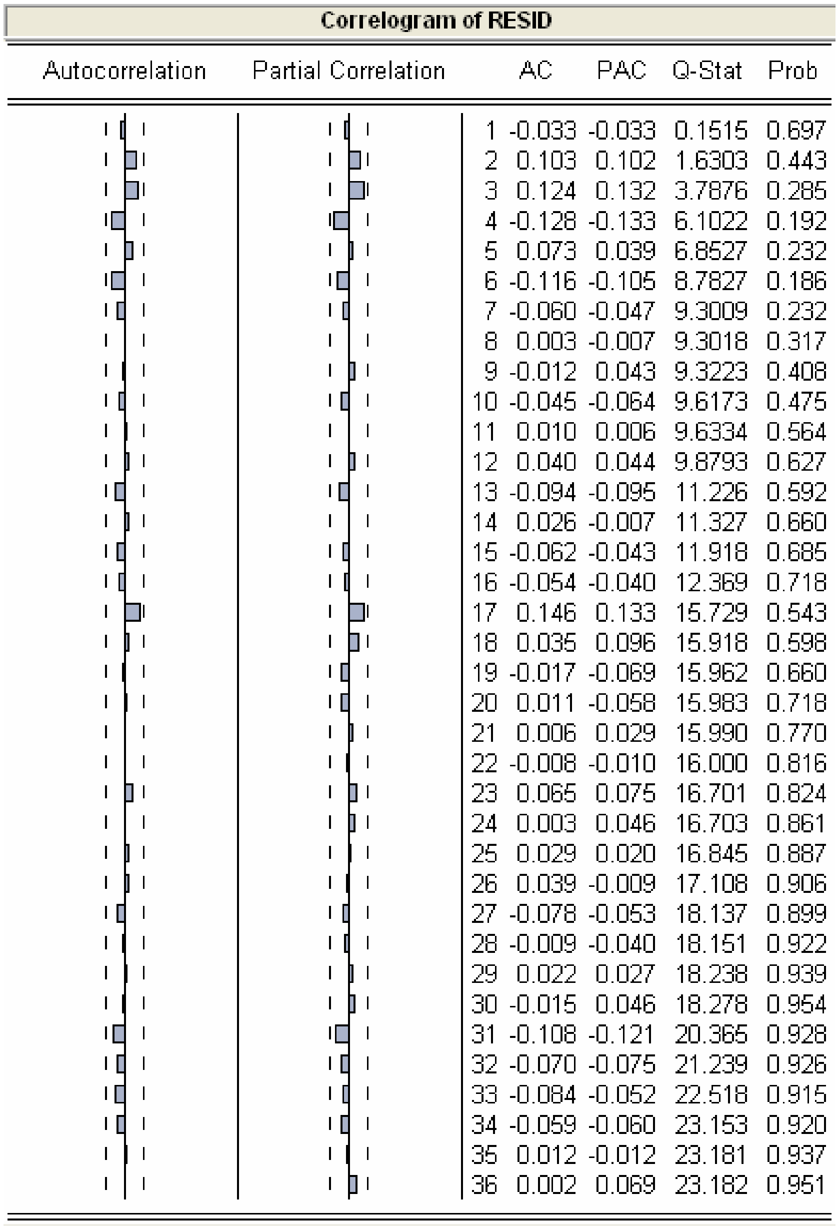
Since the seasonal differencing was chosen, then (3.14) was notated with d = 0, D = 1 and s = 12 to define z_t as:

$$\begin{aligned}z_t &=\nabla_s^D\nabla^d y_t \ =(1-B^s)^1(1-B)^0 y_t \\ &=(1-B^s)y_t \ = y_t - B^s y_t = y_t - y_{t-s} \\ &= y_t - y_{t-12}\end{aligned} \tag{4.17}$$

4.2.5.3 Diagnostic Checking

Diagnostic checking was applied to the selected model. The correlogram and residual plots are presented in Figure 4.8. The correlogram is plotted to identify the randomness. The error is found to be white noise, with the element of both SAC and SPAC falling inside the critical values. There is no serial correlation in the residual because the SAC and SPAC values at all lag are nearly zero and are within the 95% confidence interval.

Lagrange multiplier (LM) test was also applied for the first lag period (lag 1 – lag 12) and the result is presented in Table 4.7. There is no correlation up to the order 12 because the t-Statistic is less than |0.2|. This led to the conclusion that the selected model is fit.



Correlogram of RESID

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.033	-0.033	0.1515	0.697
		2	0.103	0.102	1.6303	0.443
		3	0.124	0.132	3.7876	0.285
		4	-0.128	-0.133	6.1022	0.192
		5	0.073	0.039	6.8527	0.232
		6	-0.116	-0.105	8.7827	0.186
		7	-0.060	-0.047	9.3009	0.232
		8	0.003	-0.007	9.3018	0.317
		9	-0.012	0.043	9.3223	0.408
		10	-0.045	-0.064	9.6173	0.475
		11	0.010	0.006	9.6334	0.564
		12	0.040	0.044	9.8793	0.627
		13	-0.094	-0.095	11.226	0.592
		14	0.026	-0.007	11.327	0.660
		15	-0.062	-0.043	11.918	0.685
		16	-0.054	-0.040	12.369	0.718
		17	0.146	0.133	15.729	0.543
		18	0.035	0.096	15.918	0.598
		19	-0.017	-0.069	15.962	0.660
		20	0.011	-0.058	15.983	0.718
		21	0.006	0.029	15.990	0.770
		22	-0.008	-0.010	16.000	0.816
		23	0.065	0.075	16.701	0.824
		24	0.003	0.046	16.703	0.861
		25	0.029	0.020	16.845	0.887
		26	0.039	-0.009	17.108	0.906
		27	-0.078	-0.053	18.137	0.899
		28	-0.009	-0.040	18.151	0.922
		29	0.022	0.027	18.238	0.939
		30	-0.015	0.046	18.278	0.954
		31	-0.108	-0.121	20.365	0.928
		32	-0.070	-0.075	21.239	0.926
		33	-0.084	-0.052	22.518	0.915
		34	-0.059	-0.060	23.153	0.920
		35	0.012	-0.012	23.181	0.937
		36	0.002	0.069	23.182	0.951

Figure 4.8 SAC and SPAC Correlogram of Salmonellosis Model Residual

Table 4.7 LM Test Result for Salmonellosis Residual

Variable	**Std. Error**	**t-Statistic**	**Prob.**
RESID(-1)	0.093	-0.225	0.822
RESID(-2)	0.092	1.252	0.213
RESID(-3)	0.093	1.448	0.150
RESID(-4)	0.095	-1.345	0.181
RESID(-5)	0.096	0.730	0.467
RESID(-6)	0.096	-1.141	0.256
RESID(-7)	0.098	-0.547	0.586
RESID(-8)	0.096	0.036	0.971
RESID(-9)	0.224	-0.250	0.803
RESID(-10)	0.096	-0.647	0.274
RESID(-11)	0.094	0.178	0.721
RESID(-12)	0.156	0.943	0.432

4.2.5.4 Forecasting

ARIMA(9,0,14)(12,1,24)12 has been selected as the most appropriate model from various traces. It was used to forecast the incidence from 2007 through 2009 (t_{169} – t_{204}), the results are presented in section 4.2.7.

4.2.6 Neural Network Result

A total of 168 Salmonellosis monthly data were collected. The first 12 data were used for the input. The remaining 156 data were divided into three parts: 84 data for training (54%), 36 data for selection (23%), and 36 data were used for testing (23%).

The first step in forecasting based on ANN was to initialize the network. In this experiment, the network is divided into 3 layers: input layer, hidden layer, and output layer. The hidden layer consists of 1 layer. The network was trained for 1500 iterations with 600 networks retained. The numbers of hidden nodes were varied from 1 to 12 using backpropagation algorithm.

The best network was selected from the calculation of error measures. The best model was judged by the least error measure using the mean absolute percentage error (MAPE) as the indicator value. The MAPE can be calculated using the formula:

$$MAPE\ (\%) = \frac{1}{n}\sum_{k+1}^{n}\left|y_{t+k} - y'_{t+k}\right| \tag{4.18}$$

Where y_{t+k} is the actual value and y'_{t+k} is the forecast value.

The highest performance was obtained at the iteration index 508 for the three layer network of input layer, hidden layer, and output layer consisting of 12 nodes, 5 nodes, and 1 node respectively. The network performance is presented in Table 4.8 and the network architecture is as illustrated in Figure 4.9.

Table 4.8 Salmonellosis Network Performance

Profile	**MLP s12 1:12-5-1:1**
Train Perf.	0.267
Select Perf.	0.413
Test Perf.	0.571
Train Error	0.076
Select Error	0.116
Test Error	0.162
Training/Members	BP100
Inputs	1
Hidden(1)	5
Hidden(2)	0

Table 4.8 can be described as follows. The profile indicates the network type, the number of input and output variables, the number of layers, and the number of neurons in each layer. In Statistica software, the format is <type> <inputs>:<layer1>-<layer2>-<layer3>:<outputs>, where the number of layers may vary. Then, MLP s12 1:12-5-1:1 is defined as a Multilayer Perceptron with one input variable and one output variable, and three layers consisting of input, hidden and, output with 12, 5, and 1 unit respectively, where it is clearly seen in Figure 4.9. Train perf., select perf.

and test perf. show the network performance on the training, selection, and subset test. The values of each performance indicate the ratio of predicted values to observe standard deviations. While train error/select error/test error shows error rates of the subsets. Training/members describe the training algorithm used to train the network. Referring to Table 4.8, BP100 reports one hundred epochs of back propagation.

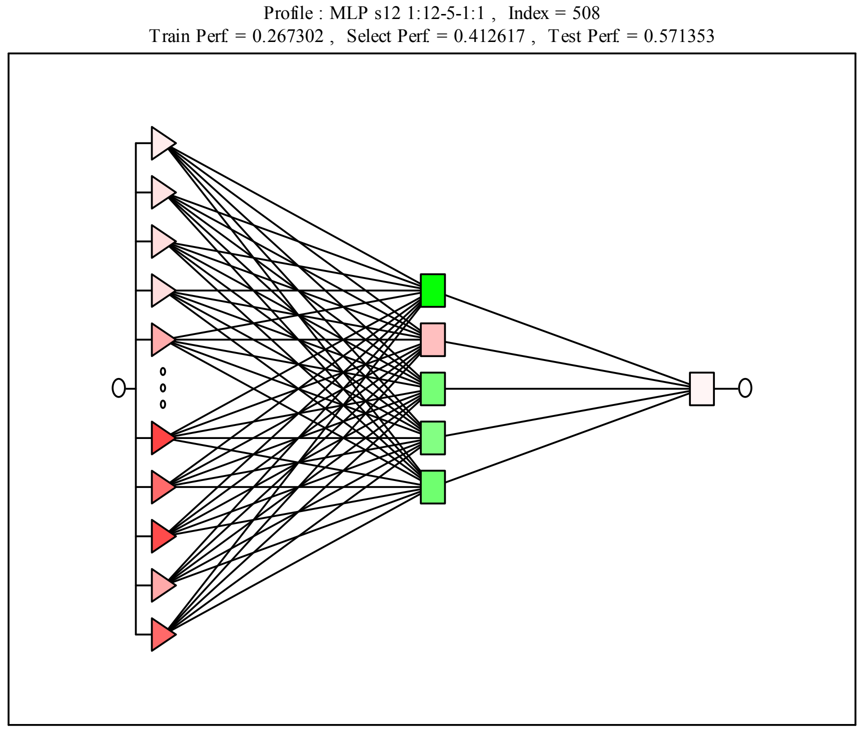

Profile : MLP s12 1:12-5-1:1 , Index = 508
Train Perf. = 0.267302 , Select Perf. = 0.412617 , Test Perf. = 0.571353

Figure 4.9 Architecture of Salmonellosis Model

4.2.7 Forecast Result

Table 4.9 shows the forecasted values for 2008 and 2009 based on the data from 1993-2006 that are generated by six different methods.

Table 4.9 Forecast Results for Salmonellosis

t	Moving Average	Linear Regression	Decomposition	Holt-Winter's	ARIMA	Neural Network
169	1908.401	1841.037	1679.528	2014.143	1678.571	1736.005
170	1955.771	1927.68	1752.651	2091.649	1965.707	2093.545
171	2328.208	2291.68	2127.18	2477.025	1969.415	2644.525
172	2546.233	2460.18	2301.582	2648.342	2692.214	2450.503
173	3081.162	2993.751	2801.836	3161.034	2657.628	2891.530
174	3814.359	3651.537	3491.539	3856.568	3364.616	3768.410
175	5158.986	4846.109	4623.556	5030.354	5020.626	4554.806
176	5590.247	5313.894	5153.509	5525.313	4675.720	5848.143
177	5431.069	5049.751	4763.859	5156.620	5655.739	5898.201
178	4780.306	4733.323	4587.943	4967.955	5691.898	3833.266
179	3878.125	3635.18	3464.678	3853.156	3489.930	3798.131
180	5907	4899.68	4629.448	5036.239	5639.062	5304.583
181	1978.441	1852.323	1594.011	2079.757	1678.558	1401.877
182	2079.143	1938.966	1667.134	2157.263	1965.666	2405.847
183	2488.139	2302.966	2041.664	2542.639	1969.399	2678.292
184	2683.636	2471.466	2216.065	2713.956	2692.24	2114.184
185	3288.235	3005.037	2716.319	3226.648	2657.604	3389.266
186	4039.041	3662.823	3406.022	3922.182	3364.646	3436.217
187	5403.701	4857.395	4538.039	5095.968	5020.595	4856.183
188	5967.822	5325.18	5067.992	5590.927	4675.72	6201.189
189	5719.066	5061.037	4678.342	5222.234	5655.587	5051.095
190	5423.863	4744.609	4502.426	5033.57	5691.944	4106.460
191	4246.469	3646.466	3379.161	3918.77	3489.946	3912.961
192	5907	4910.966	4543.932	5101.853	5639.14	5064.932

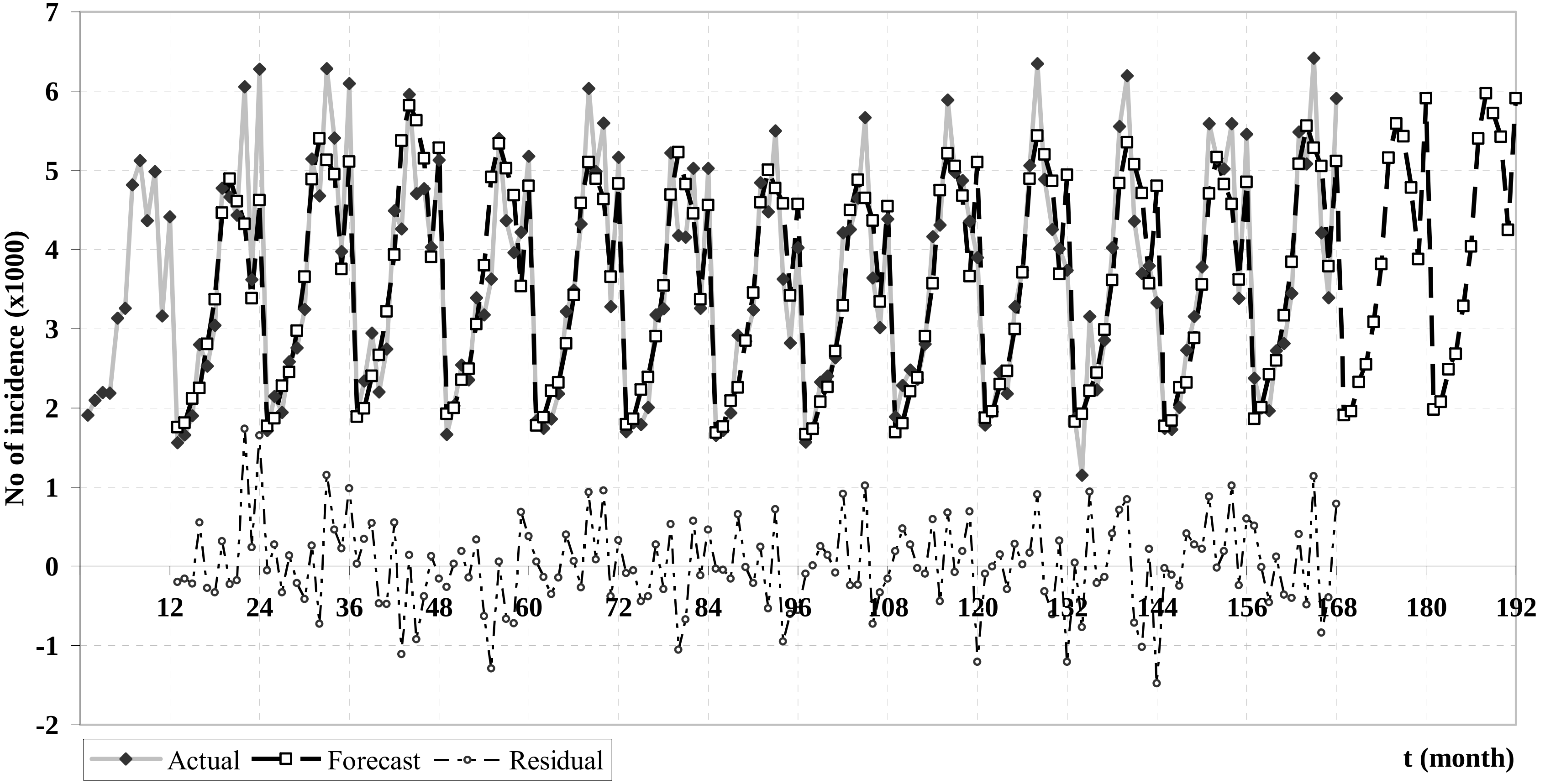

No of incidence (x1000)
7
6
5
4
3
2
1
0
-1
-2
12
24
36
48
60
72
84
96
108
120
132
144
156
168
180
192
t (month)
Actual
Forecast
Residual

Figure 4.10 Moving Average Result of Salmonellosis

Figure 4.10 shows a sample of completed Salmonellosis time series plots of moving average method. This chart consists of three components: actual values, fitted values, and residual values that represented by grey solid line, the black dashed line, and the dotted black line, respectively. The 156 monthly forecasts (January 1994 – December 2006) generated using the moving average model is compared with the actual data. The forecast errors for the entire month are presented as the residual values. The residual values are resulted from the differences between actual data and forecasted value in associated month. Figure 4.10 also provides the predicted number of incidence in 2008 and 2009. The rest of the forecast plots of regression, decomposition, Holt-Winter's, ARIMA, and neural network are given in Figure B1.1, B1.2, B1.3, B1.4, and B1.5 of Appendix B.

4.2.8 Analysis of Variance (ANOVA)

There were 6 different techniques, namely moving average, linear regression, decomposition, Holt-Winter's, ARIMA, and neural network. The forecast values of all methods were compared with actual data using ANOVA.

Let $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6, \mu_7$ were the average values obtained from actual data, followed by the average estimation results from moving average, linear regression, decomposition, Holt-Winter's, ARIMA, and neural network, respectively. The hypothesis was:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$$

Where the means of all groups of forecast results are equal.

$$H_1 : \mu_i \neq \mu_j \quad i, j = 1,2,3,4,5,6,7, i \neq j$$

Where not all the means are equal.

Table 4.10 Blocked Design ANOVA for Salmonellosis

SUMMARY

Groups	Count	Sum	Average	Variance
Actual Data	135	488458	3618.207	1829864
Moving Average	135	488826.8	3620.939	1618502
Linear Regression	135	488677.6	3619.834	1564833
Decomposition	135	503236.4	3727.677	1541855
Holt-Winter's	135	489667.9	3627.169	1586876
ARIMA	135	491894.9	3643.666	1792091
Neural Network	135	487275.1	3609.445	1763460

Source of Variation	SS	df	MS	F	P	F_{crit}(α=0.05)
Between Group (Method)	1352625	6	225437.5	2.709851	0.013038	2.109839
Blocks (month)	1.5E+09	134	11198331	134.6085	0	1.231085
Within Groups (Error)	66886244	804	83191.85			
Total	1.57E+09	944				

The hypothesis was tested by using an assumed level of significance (α) equal to 0.05. The ANOVA results for this case study are shown in Table 4.10. Since the difference of the method also yields the different starting point of forecast results, for example the forecast value for ARIMA could be predicted in t = 34. Thus, Balance ANOVA was used where actual data and all forecast values started from t = 34 and it resulted 135 data as seen in Table 4.10. Refer to the Table 4.10, if the *F* value of the method is greater than the corresponding value of F_{crit}, then it is concluded that the result predicted by each method is significantly different. Therefore the null hypothesis is rejected. Otherwise, if the F value is less than F_{crit}, then it is concluded that the predicted result is not significantly different.

The ANOVA table is divided into 'between group effects' (effects due to experiment) and 'within group effects' (the unsystematic variation/error in the data). The experiment has a variation in the results which arise from a systematically controlled source. Time (months) is the common source of variation that can be controlled by blocking. Then, a blocked design of ANOVA was applied with month as the blocking parameter.

Table 4.10 presents statistical information obtained by ANOVA.

- The first column shows the source of variability (variability between column is represented by between group/method, variability between rows is represented by months as the blocking parameter, and variability within groups is represented by the error).
- The second column shows the total experimental effect as the sum of square (SS).
- The third column shows the degree of freedom (*df*) associated with each source.
- The fourth column presents the average of experimental effect as the mean square (MS), calculated from the ratio of SS/*df* .
- The fifth column presents the *F* value from the ANOVA mathematical calculation associated with the source of variation.
- The sixth column shows the probability that the variations between results are random. If the *P* values < 0.05 then the null hypothesis is rejected.

- And the seventh column presents the critical values from the f-distribution in the statistical table based on two values of degrees of freedom (F_{crit}).

From the ANOVA result shown in Table 4.10, SS (α= 0.05) is 66886244 and MS is 83191.85. It also can be concluded that at $\alpha = 0.05$, the null hypothesis is rejected because the value of F is larger than the value of F_{crit}, where $F_{crit} = 2.109839$ and $F = 2.709851$.

The P value of 'between group' also give an evidence to reject the null hypothesis, where $P = 0.013038 < 0.05$.

The blocked ANOVA tested whether any of the population means differ from each other. While, a multiple comparison test was conducted to check which population means differ from which others mean. The following section reports the application of Duncan's Multiple Range Test.

4.2.9 Duncan Multiple Range Test

The Duncan multiple range tests are used to test which methods having different mean.

Let μ_1 represent the mean of the actual data.

Let μ_2 represent the mean of moving average result.

Let μ_3 represent the mean of regression analysis result.

Let μ_4 represent the mean of decomposition result.

Let μ_5 represent the mean of Holt-Winter's result.

Let μ_6 represent the mean of ARIMA analysis result.

Let μ_7 represent the mean of neural network result.

For the confidence level $\alpha = 0.05$, the least significant range (LSR) was found from the Duncan's Table. The LSR was computed as shown below:

Mean sum of square error, MS = 83191.85, n = 135

Then, standard error of each average, $S = \sqrt{\dfrac{MS}{n}}$ (4.19)

$$S = 24.824$$

Since the total number of selected groups was 7, the total number of ranges equal to 6. The values of p are calculated as the range difference, where they are obtained from the subtraction of mean ranks and added by one. As the example: sample 1 and sample 2, then p = (2-1)+1=2.

From the table of significant ranges Montgomery, for 804 degrees of freedom (804 is the number of degrees of freedom for within groups from ANOVA table or total number of samples) and α = 0.05, the six ranges were calculated as given below:

$$r_2 = r_{\alpha(p,df)} = r_{0.05(2,804)} = 2.772$$

$$r_3 = r_{\alpha(p,df)} = r_{0.05(3,804)} = 2.918$$

$$r_4 = r_{\alpha(p,df)} = r_{0.05(4,804)} = 3.017$$

$$r_5 = r_{\alpha(p,df)} = r_{0.05(5,804)} = 3.089$$

$$r_6 = r_{\alpha(p,df)} = r_{0.05(6,804)} = 3.146$$

$$r_7 = r_{\alpha(p,df)} = r_{0.05(7,804)} = 3.193$$

LSR can be calculated from the equation below:

$$R_p = r_p \times S \qquad (4.20)$$

By applying (4.20), LSR results were:

$$R_2 = r_2 \times S = r_{0.05(2,804)} \times S = 2.772 \times 24.824 = 68.812$$

$$R_3 = r_3 \times S = r_{0.05(3,804)} \times S = 2.918 \times 24.824 = 72.437$$

$$R_4 = r_4 \times S = r_{0.05(4,804)} \times S = 3.017 \times 24.824 = 74.894$$

$$R_5 = r_5 \times S = r_{0.05(5,804)} \times S = 3.089 \times 24.824 = 76.682$$

$$R_6 = r_6 \times S = r_{0.05(6,804)} \times S = 3.146 \times 24.8240 = 78.097$$

$$R_7 = r_7 \times S = r_{0.05(7,804)} \times S = 3.193 \times 24.824 = 79.263$$

A mean symbol to each method and the values was assigned as shown in Table 4.11 below.

Table 4.11 Mean Symbol of Each Group

Methods	Mean Symbol	Mean Value
Actual	M1	3618.207
Moving Average	M2	3620.939
Regression	M3	3619.834
Decomposition	M4	3727.677
Holt-Winter's	M5	3627.169
ARIMA	M6	3643.666
Neural Network	M7	3609.445

The mean values in Table 4.11 were sorted and tabulated as in Table 4.12.

Table 4.12 The Sorted Mean Values of Salmonellosis

Method	**M7**	**M1**	**M3**	**M2**	**M5**	**M6**	**M4**
Mean	3609.445	3618.207	3619.834	3620.939	3627.169	3643.666	3727.68

The differences between every mean value and all other mean values were calculated and the results are listed below:

M1 – M2 = 2.732 < 68.812 (R_2)

M1 – M3 = 1.627 < 72.437 (R_3)

M1 – M4 = 109.470 > 74.894 (R_4)

M1 – M5 = 8.962 < 76.682 (R_5)

M1 – M6 = 25.459 < 78.097 (R_6)

M1 – M7 = 8.762 < 79.263 (R_7)

M2 – M3 = 1.105 < 68.812 (R_2)

M2 – M4 = 106.738 > 72.437 (R_3)

M2 – M5 = 6.230 < 74.894 (R_4)

M2 – M6 = 22.727 < 76.682 (R_5)

M2	–	M7	=	11.494	<	78.097	(R_6)
M3	–	M4	=	107.843	>	68.812	(R_2)
M3	–	M5	=	7.335	<	72.437	(R_3)
M3	–	M6	=	23.832	<	74.894	(R_4)
M3	–	M7	=	10.389	<	76.682	(R_5)
M4	–	M5	=	100.508	>	68.812	(R_2)
M4	–	M6	=	84.011	>	72.437	(R_3)
M4	–	M7	=	118.232	>	74.894	(R_4)
M5	–	M6	=	16.497	<	68.812	(R_2)
M5	–	M7	=	17.724	<	72.437	(R_3)
M6	–	M7	=	34.221	<	68.812	(R_2)

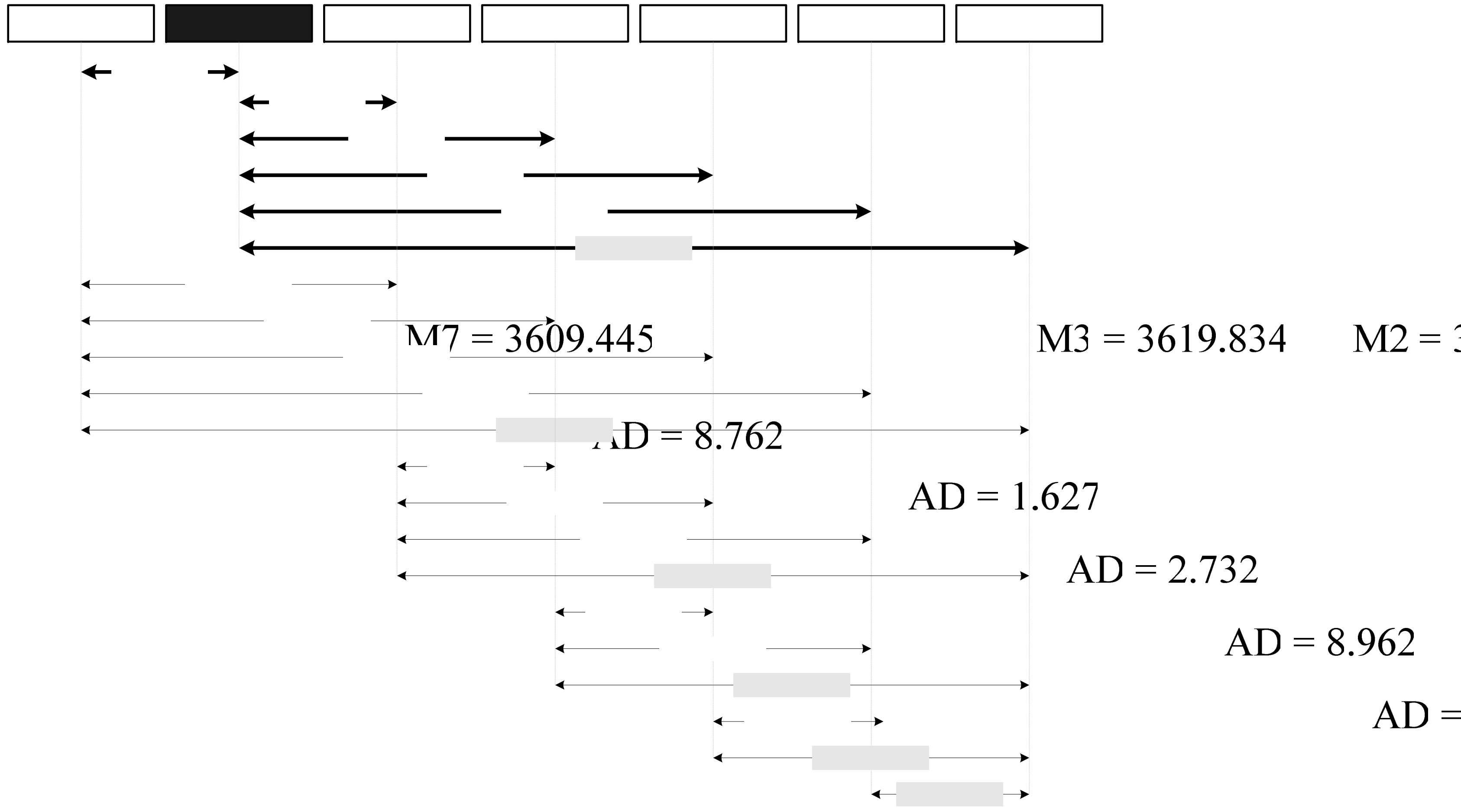


Figure 4.11 Actual Differences between Different Pairs of Means in Comparing Different Forecasting Methods

In order to distinguish the significant difference between group means from various forecasting methods, the AD for the associated pair is formatted by using bold font and grey background as shown in Figure 4.11. The first six AD show the differences between actual data and the forecast results from the different methods. The difference between mean value of the decomposition method (M4 = 3727.677) with the other five methods is large, and also the AD is greater than its least significant range (109.470 > 74.894). The decomposition method also performs badly than the other techniques. This is evident from the difference between the mean of the decomposition method with that of other methods, which is greater from the least significant range (M2 – M4 = 106.738 > 72.437, M3 – M4 = 107.843 > 68.812, M4 – M5 = 100.508 > 68.812, M4 – M6 = 84.011 > 72.437, M4 – M7 = 118.232 > 74.894). This indicates that the decomposition method is not suitable to be applied to the historical data.

On the contrary, there is no difference in the mean comparison between actual data and other methods (regression, moving average, Holt-Winter's, ARIMA, neural network). Since there is no significant difference between actual data and the rest of the methods, it is concluded that the regression, moving average, Holt-Winter's, ARIMA, and neural network could be used mainly for predicting future number of Salmonellosis incidence in human within the same historical data.

4.2.10 Performance of Forecast Results

In order to determine the most appropriate methods for the Salmonellosis dataset, the suitable methods (moving average, regression, Holt-Winter's, ARIMA, neural network) was ranked by using coefficient of variation (CV) values. Table 4.13 gives the output of CV values from the forecast results.

Table 4.13 shows the standard deviation, the mean, the coefficient of variation (CV) and rank of each method. The method ranks reported in the table are based on the value of CV. To rank these five methods, the mean and standard deviation are firstly computed for each of them. Then, CV is used as the basis of comparison between mean and standard deviation. The CV of each method is calculated as standard deviation divided by mean. A high CV value reflects inconsistency among the samples within the group of forecast results [161]. Thus, method with the smallest CV is identified as the top ranking and method with the biggest CV is determined as the lowest rank. According to Table 4.13, the order of rank from the top to the lowest are regression analysis, Holt-Winter's, moving average, ARIMA, neural network. The result suggests that regression analysis is the most appropriate method to be applied into the Salmonellosis dataset in this case study.

Table 4.13 CV Values of Forecast Results

Method	**Standard Deviation (σ)**	**Mean (μ)**	**Coefficient of Variation (CV)**	**Rank**
Moving Average	1272.203	3620.939	0.351	3
Regression Analysis	1250.933	3619.834	0.346	1
Holt-Winter's	1259.713	3627.169	0.347	2
ARIMA	1338.690	3643.666	0.367	4
Neural Network	1327.953	3609.445	0.368	5

4.3 Dialog Generation and Management Subsystem

This section discusses the results of 'what if analysis' for Salmonellosis dataset. Since the similarity of Graphical User Interface (GUI) design between both case studies, then the detailed descritiption of user interface for the proposed framework is presented in Chapter 6. The 'what if analysis' (sensitivity analysis) is used to assess the performance of the output variable as a result of the variation in the input variable (called as scenario). It simulates the system to find as many possible outputs from some probable scenario. This analysis is able to identify the fluctuation in the forecasting results in a specific method that is influenced by changes in the historical data.

The 'what if analysis' is performed according to the four steps outlined below:

- *Step 1: Identify key variables*

  The aim of this step is to understand what kind of information is available to drive the process. Since the zoonosis forecasting model is based on the collection of monthly data, then monthly data of incidence were used as the key variable.

- Step 2: Generate what-if question

  Hypothetical postulate is generated in this step. The what-if query in the zoonosis prediction is:

  > 'How would the forecast values change if more historical data are added?'

  In the forecasting development model, 168 monthly data from 1993 – 2006 were used. Then, twelve monthly data (from 2007) were added to the previous dataset to obtain the future values of incidence. Based on the what-if query, the following scenarios were constructed:

  Scenario 1: Forecast number of incidence in 2008 and 2009 based on the historical data from 1993 – 2006 (168 months).

  Scenario 2: Forecast number of incidence in 2008 and 2009 based on historical data from 1993 – 2007 (180 months).

- Step 3: Calculate the effect of changes in the variables

  The forecast values should be recalculated for different values of key variables; and the result was used to calculate the percentage of sensitivity using the formula (3.32).

  The sensitivity formula was applied repeatedly to all methods (regression analysis, moving average, decomposition, Holt-Winter's, ARIMA, and neural network).

- Step 4: Analyze the scale of changes

  In this step, the percentage of sensitivity for 2008 and 2009 in corresponding months were compared and analyzed.

Table 4.14 summarizes the sensitivity analysis for moving average method. The table is separated into two groups. The first group presents the percentage of sensitivity analysis in the forecasted number of incidence for 2008. The percentage was obtained from the difference between the forecasted results based on 1993–2006 data and 1993–2007 data as mentioned in Equation (3.33). The second group shows the sensitivity analysis of the forecasted number of incidence in 2009 that have been obtained by using the same formula in the forecasted incidence in 2008. The positive sensitivity indicates that the forecasted values based on the data from 1993–2007 is higher than those that resulted from 1993– 2006 data. The negative sign shows that the forecasted results based on data from 1993– 2007 is lower than those from 1993–2006. The sensitivity analysis of other methods is given in Appendix C, where Table C1.1 for regression, Table C1.2 for decomposition, Table C1.3 for Holt-Winter's, Table C1.4 for ARIMA and Table C1.5 for neural network.

Table 4.14 Sensitivity of Moving Average Salmonellosis

Month	**Forecast 2008**			**Forecast 2009**		
	Based data 1993-2006	**Based data 1993-2007**	**Sensitivity (%)**	**Based data 1993-2006**	**Based data 1993-2007**	**Sensitivity (%)**
Jan.	1921.746	2082.300	7.71%	1978.441	2016.362	1.88%
Feb.	2015.587	2090.444	3.58%	2079.143	2066.84	-0.60%
Mar.	2411.867	2495.847	3.36%	2488.139	2476.524	-0.47%
Apr.	2602.532	2607.976	0.21%	2683.636	2620.794	-2.40%
May	3185.448	3213.201	0.86%	3288.235	3209.435	-2.46%
Jun.	3913.608	3997.683	2.10%	4039.041	3977.262	-1.55%
Jul.	5230.943	5100.773	-2.55%	5403.701	5192.54	-4.07%
Aug.	5756.117	5667.135	-1.57%	5967.822	5707.182	-4.57%
Sept.	5507.085	5643.754	2.42%	5719.066	5540.331	-3.23%
Oct.	5182.820	5025.129	-3.14%	5423.863	5099.708	-6.36%
Nov.	4123.687	3800.013	-8.52%	4246.469	3911.02	-8.58%
Dec.	5907.000	5486.000	-7.67%	5907	5486	-7.67%

The moving average monthly forecast for 2008 is plotted as in Figure 4.12, where two sets of values are shown; one based on the data from 1993 – 2006 and the other based on the data from 1993 – 2007. The 2009 forecast results for the two datasets are plotted as in Figure 4.13. The plots of sensitivity of the regression, decomposition, Holt-Winter's, ARIMA, and neural network can be found in Table D1.1, Table D1.2, Table D1.3, Table D1.4, and Table D1.5 of Appendix D.

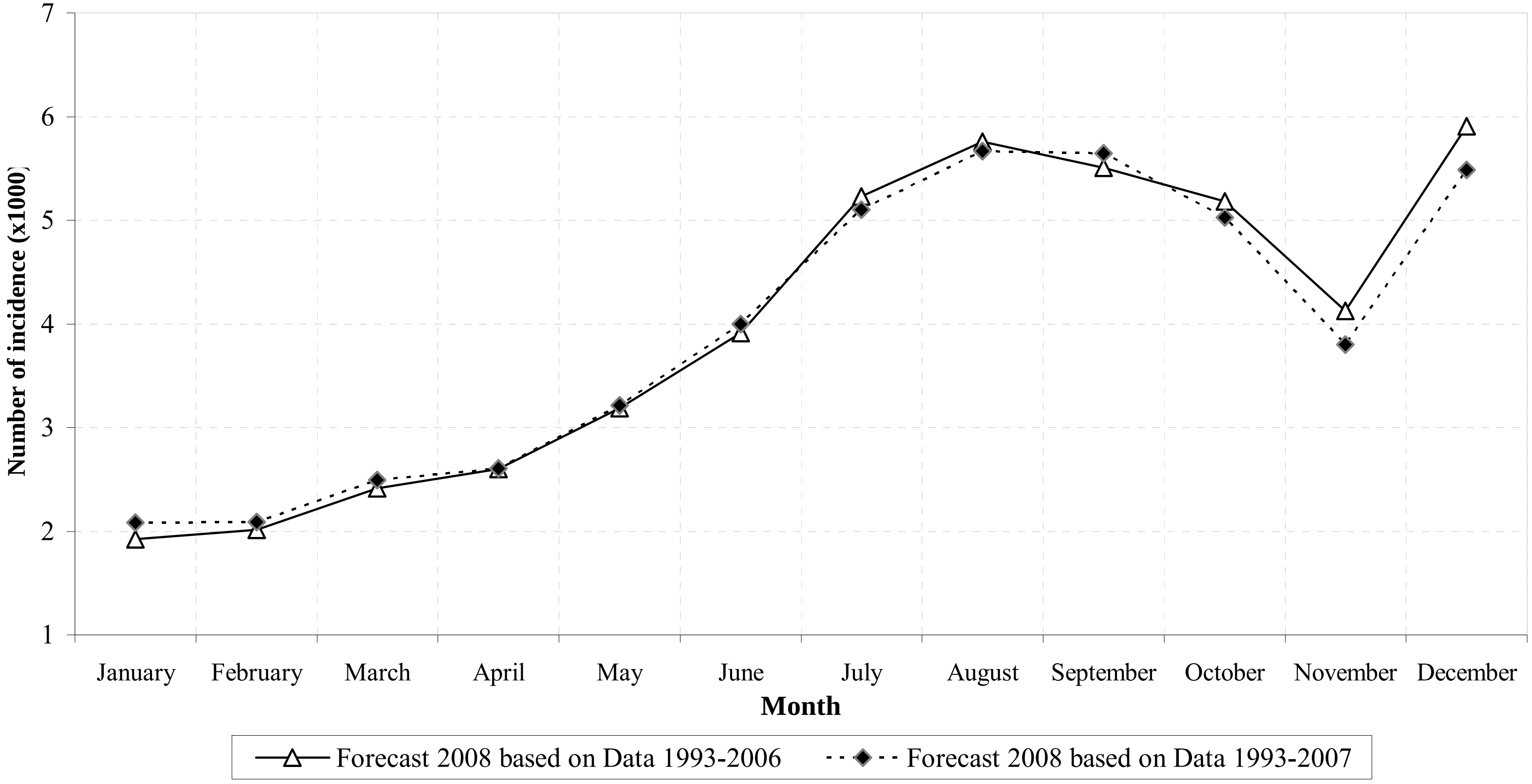


Figure 4.12 Sensitivity Analysis of Moving Average 2008 Forecast for Salmonellosis

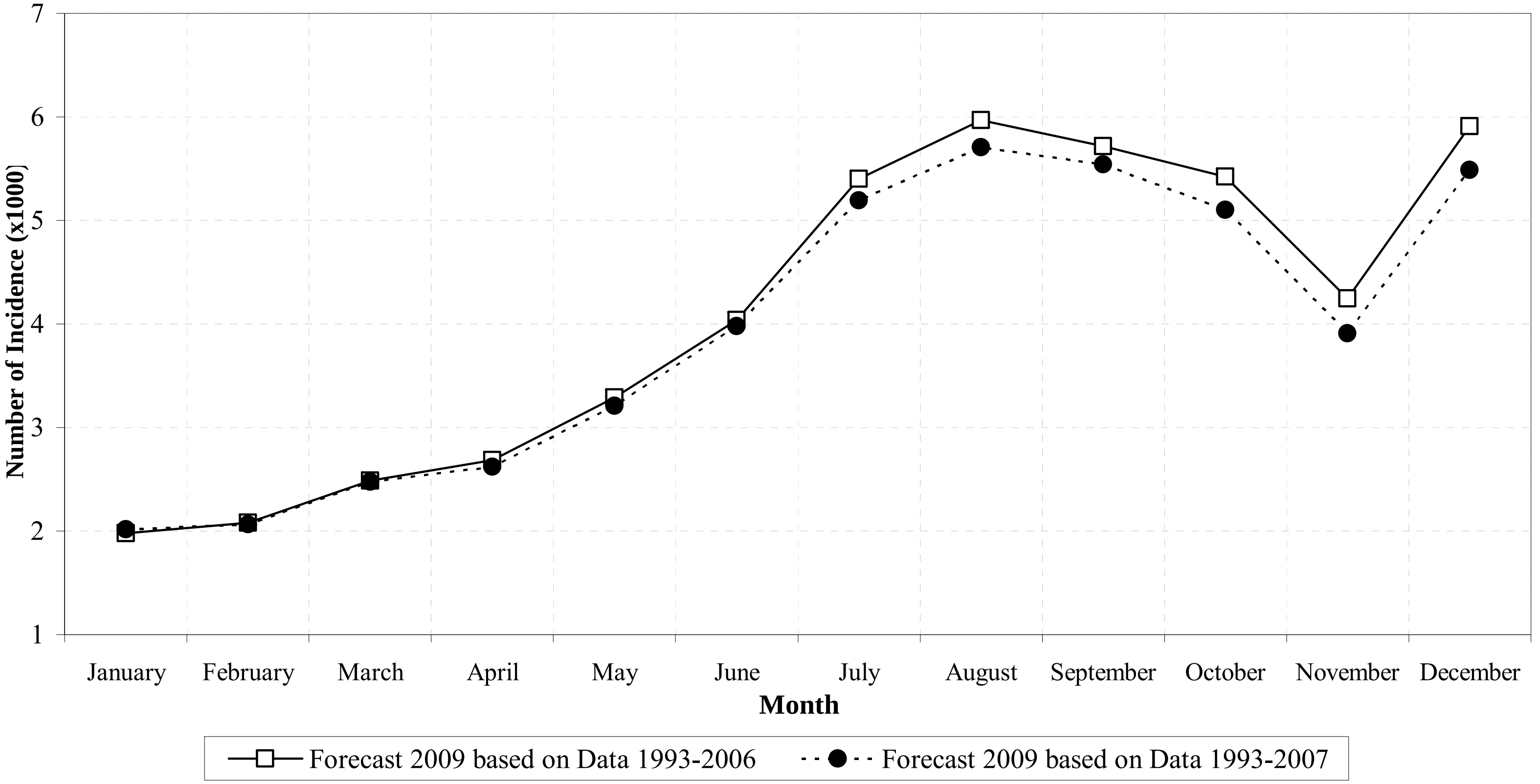


Figure 4.13 Sensitivity Analysis of Moving Average 2009 Forecast for Salmonellosis

4.4 Summary

The thesis covered two types of case study, time series with additive seasonal trend and also time series with multiplicative seasonal trend. This chapter focuses on the development of a DSS framework based on additive seasonal. Salmonellosis has been chosen for this case study because the data exhibited a relatively constant trend. The DSS framework was applied to two sets of Salmonellosis data, from 1993 – 2006 and from 1993 – 2007. Hence, the three components of the DSS framework have been fully described in this chapter. The results have shown that this case study has helped to provide a better proof on how to perform and apply an additive seasonal model in the framework into a real case. In Chapter 5, the results of the framework based on multiplicative seasonal time series are reported.

CHAPTER 5
DSS FRAMEWORK (CASE STUDY ON TUBERCULOSIS)

5.0 Chapter Overview

Having continuity with Chapter 4, this chapter focuses on the reporting of DSS result in Tuberculosis. Section 5.1 describes the result of data management subsystem. This is followed by Section 5.2 discussing the overall result of model management subsystem. Next, Section 5.3 reports dialog generation and management subsystem, especially related to What-If analysis results. Lastly, a summary of the chapter is presented in Section 5.4.

5.1 Data Management Subsystem

The time series was collected from the number of Tuberculosis incidence in human from Morbidity and Mortality Weekly Report (MMWR) published by the Center for Disease Control and Prevention (CDC) of the United State for the 168 month period from January 1993 to December 2006. The seasonal variation of the original data is presented in Figure 5.1.

Figure 5.2 shows the seasonal stacked line for human Tuberculosis incidence in US from 1993 until 2006. The figure shows time series seasonal factor as derived from Figure 5.1. For each group, the mean of the data from a specific month are illustrated by horizontal lines. Thus, it can be defined that a peak season of incidence occurs in December and the minimum number of incidence occurs in January. A time series plot of the historical data shows the seasonal variations, which exhibits similar trend every year.

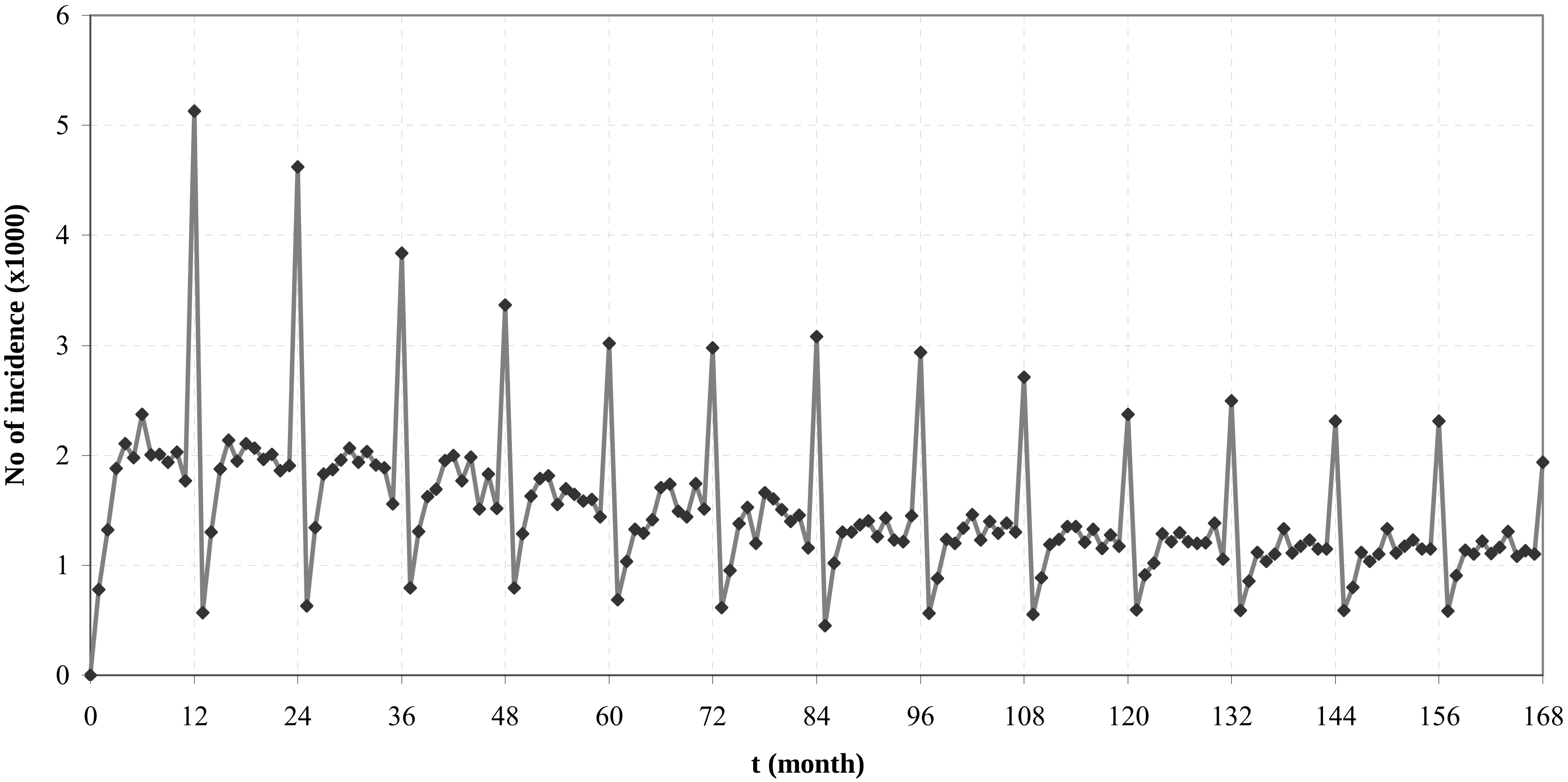


Figure 5.1 Monthly Number of Tuberculosis Incidence in US (1993-2006)

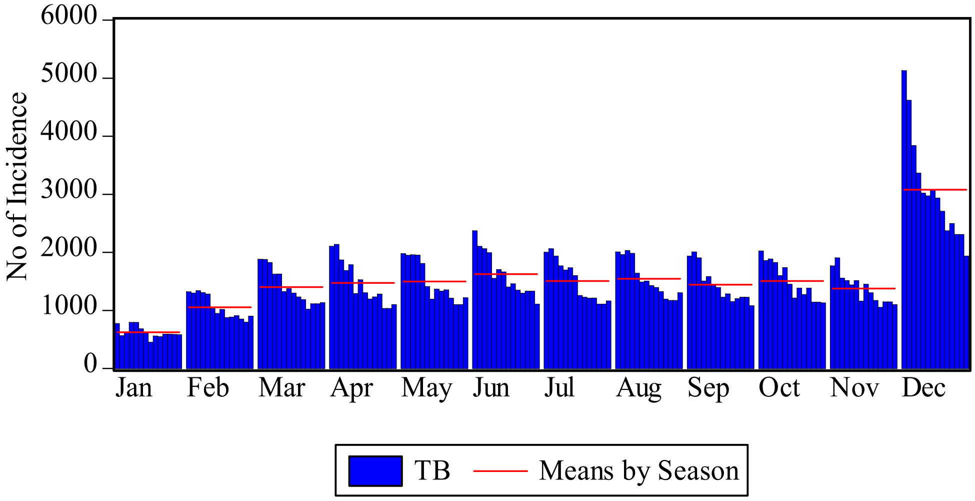


Figure 5.2 Seasonal Stacked Line of US Tuberculosis (1993-2006)

Upon further analysis of time series, it was observed that the number of incidence in any month of the year remain significantly decreasing. Thus, information about time series trend could be used to determine the suitable approach of forecasting monthly diseases incidence in the model management subsystem.

5.2 Model Management Subsystem

Six forecasting methods were applied in this model management subsystem. The results are presented in following subsections.

5.2.1 Moving Average Result

Firstly, the seasonal factor needed to be calculated. The moving average process used twelve historical data where it was started with January as the first month of group and December as the twelfth month of group. Seasonal factors were obtained by dividing the average between annual group mean by the average of the whole data, with the following question. For example, the sequences Tuberculosis data of January 1993 to January 2006 were 778, 567, 632, 794, 794, 685, 613, 453, 563, 552, 593, 591, 589, and 583. Thus, the average of January was 627.643. For the Tuberculosis, the average of the 168 data from January 1993 to December 2006 was 1513.634, thus the seasonal factor of January was 0.415.

Table 5.1 presents seasonal factors calculation of moving average using Tuberculosis time series. The values of the seasonal factor represent the level of incidence in every month. According to Table 5.1, the peak of incidence occurs in December with the seasonal factor 2.038 and the lowest incidence occurs in January with the seasonal factor 0.415.

After obtaining the seasonal factors, the existing dataset need to be deseasonalized first by dividing the actual data with the seasonal factor. In this research, the number of terms in each moving average was 12. In order to calculate moving average, the forecasted results were calculated from the average of the preceding twelve data. It led to the result at the thirteenth step of forecasting model. Finally, the results of the averaging process were multiplied by the seasonal factor to get the final prediction values. The forecast for 2008 and 2009 are presented in Section 5.2.7.

Table 5.1 Moving Average Seasonal Factor for Tuberculosis

Month	Seasonal Factor
January	0.415
February	0.699
March	0.925
April	0.978
May	0.992
June	1.073
July	0.998
August	1.022
September	0.953
October	1.000
November	0.906
December	2.038

5.2.2 Regression Analysis Result

The time series used 12 seasonal components and 12 ordinal variables were defined. A month corresponding to the actual values was assigned as 1 while the rest are assigned into 0. Application of formula (3.3) resulted in the level, trend, and seasonality variables presented Table 5.2.

Table 5.2 Regression Coefficients for Tuberculosis

Variable	Coefficient
Level	3653.683
Time	-6.38855
January	-2521.35
February	-2085.06
March	-1732.35
April	-1657.68
May	-1626.22
June	-1491.9
July	-1603.23
August	-1558.34
September	-1654.52
October	-1585.71
November	-1709.82

The regression formula is:

$$\begin{aligned} y_t &= TR_t + SN_t + \varepsilon_t \\ &= \beta_0 + \beta_1 t + \beta_{s1} x_{s1,t} + \beta_{s2} x_{s2,t} + \beta_{s3} x_{s3,t} + \beta_{s4} x_{s4,t} + \beta_{s5} x_{s5,t} \\ &\quad + \beta_{s6} x_{s6,t} + \beta_{s7} x_{s7,t} + \beta_{s8} x_{s8,t} + \beta_{s9} x_{s9,t} + \beta_{s10} x_{s10,t} + \beta_{s11} x_{s11,t} \\ &= 3653.683 - 6.3886t - 2521.35 x_{s1,t} - 2085.35 x_{s2,t} - 1732.35 x_{s3,t} \\ &\quad - 1657.68 x_{s4,t} - 1626.22 x_{s5,t} - 1491.9 x_{s6,t} - 1603.23 x_{s7,t} - 1558.34 x_{s8,t} \\ &\quad - 1654.52 x_{s9,t} - 1585.71 x_{s10,t} - 1709.82 x_{s11,t} \end{aligned} \tag{5.1}$$

Applying the use of dummy variables as in Table 4.2, the regression coefficients yield the forecasted results. For example to calculate the Tuberculosis predicted value of February 2003 ($t = 2$) is given below:

$$
\begin{aligned}
y_t &= TR_t + SN_t + \varepsilon_t \\
&= \beta_0 + \beta_1 t + \beta_{s1}x_{s1,t} + \beta_{s2}x_{s2,t} + \beta_{s3}x_{s3,t} + \beta_{s4}x_{s4,t} + \beta_{s5}x_{s5,t} \\
&\quad + \beta_{s6}x_{s6,t} + \beta_{s7}x_{s7,t} + \beta_{s8}x_{s8,t} + \beta_{s9}x_{s9,t} + \beta_{s10}x_{s10,t} + \beta_{s11}x_{s11,t} \\
&= 3653.683 - (6.3886 \times 2) - (2521.35 \times 0) - (2085.35 \times 1) - (1732.35 \times 0) \\
&\quad - (1657.68 \times 0) - (1626.22 \times 0) - (1491.9 \times 0) - (1603.23 \times 0) - (1558.34 \times 0) \\
&\quad - (1654.52 \times 0) - (1585.71 \times 0) - (1709.82 \times 0) \\
&= 1555.842
\end{aligned}
$$

Thus, the predicted result of February 2003 (t = 2) is 1555.842. The same assumption can be used to obtain the rest predicted values. It is seen from the calculation, that the seasonal factor of the connected month is multiplied by 1, while the entire seasonal factor is multiplied by 0. The results plot of forecasted values up to 2 year ahead (2008 and 2009) can be seen in Appendix B, while the forecasted values for 2008 and 2009 is presented in Section 5.2.7.

5.2.3 Decomposition Result

Since Tuberculosis time series (Figure 5.1) exhibits decreasing seasonal variation, multiplicative decomposition was used [104]. The multiplicative decomposition model is

$$y_t = TR_t \times SN_t \times CL_t \times IR_t \tag{5.2}$$

In this case, the component CL_t could be omitted from equation (5.2) because there is no cyclical component in the data. Equation (5.2) can be rewritten as,

$$y_t = TR_t \times SN_t \times IR_t \tag{5.3}$$

Where y_t is the observed value of the time series in time period t; TR_t is the trend component (or factor) in time period t; SN_t is the seasonal component (or factor) in time period t; CL_t is the cyclical component (or factor) in time period t; and IR_t is the irregular component (or factor) in time period t.

The first step in modeling decomposition method is by building a moving average (MA) and a centered moving average (CMA_t) of the time series in the decomposition method. Using the MA values, the CMA can be obtained by computing the average of the two adjacent MA values. Equation (5.3) can be rewritten as

$$SN_t \times IR_t = y_t / TR_t \tag{5.4}$$

Following the estimation of $SN_t \times IR_t$ is sn_t / ir_t then

$$sn_t \times ir_t = y_t / tr_t = y_t / CMA_t \tag{5.5}$$

The number of $\overline{sn_t}$ could be obtained by grouping the values of $sn_t \times ir_t$ by months and then calculating average, $\overline{sn_t}$, for each month. With L = 12 (number of period a year) then the seasonal factor is:

$$sn_t = \left(L / \sum_{t=1}^{L} \overline{sn_t} \right) \times \overline{sn_t} = 1.000701\left(\overline{sn_t}\right) \tag{5.6}$$

Table 5.3 shows the calculation for monthly seasonal factor presented using a graph Figure 5.3. Table 5.3 and Figure 5.3 are created based on calculation on equation (5.6). It is seen that the peak of incidence occurs in December, which is indicated by the highest seasonal factor (2.071). On the contrary, the lowest seasonal factor in January signifies the lowest number of incidence with the seasonal factor 0.419.

Table 5.3 Monthly Seasonal Factor for Tuberculosis Decomposition

Month	***sn_t x ir_t = yt/tr_t***													***sn_t***
	Year 1	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Year 6**	**Year 7**	**Year 8**	**Year 9**	**Year 10**	**Year 11**	**Year 12**	**Year 13**	
Jan	0.274	0.315	0.424	0.458	0.445	0.405	0.322	0.417	0.417	0.480	0.488	0.501	0.493	0.419
Feb	0.629	0.671	0.703	0.749	0.675	0.631	0.734	0.655	0.672	0.741	0.709	0.680	0.760	0.693
Mar	0.906	0.913	0.881	0.956	0.873	0.913	0.943	0.915	0.906	0.832	0.925	0.950	0.956	0.914
Apr	1.034	0.936	0.926	1.054	0.850	1.024	0.957	0.885	0.949	1.040	0.865	0.882	0.931	0.949
May	0.943	0.986	1.073	1.075	0.923	0.816	1.004	0.984	1.050	0.984	0.926	0.939	1.033	0.980
Jun	1.026	1.066	1.111	0.931	1.114	1.141	1.025	1.089	1.065	1.050	1.123	1.136	0.953	1.065
July	0.954	1.015	1.013	0.995	1.029	1.137	1.102	0.919	0.922	0.963	0.981	0.941	0.945	0.994
Aug	0.961	0.964	1.062	1.116	1.006	0.981	1.039	1.047	1.050	1.052	0.968	0.997	0.996	1.019
Sep	0.927	0.986	1.001	0.849	0.982	0.949	0.965	0.905	0.970	0.918	0.970	1.048	1.040	0.963
Oct	0.970	0.919	0.997	1.027	1.014	1.137	1.011	0.898	1.041	1.021	1.124	0.976	0.965	1.008
Nov	0.846	0.947	0.828	0.853	0.936	0.988	0.808	1.077	0.979	0.941	0.868	0.979	0.962	0.925
Dec	2.468	2.297	2.042	1.917	1.972	1.957	2.151	2.180	2.042	1.916	2.056	1.966	1.939	2.071

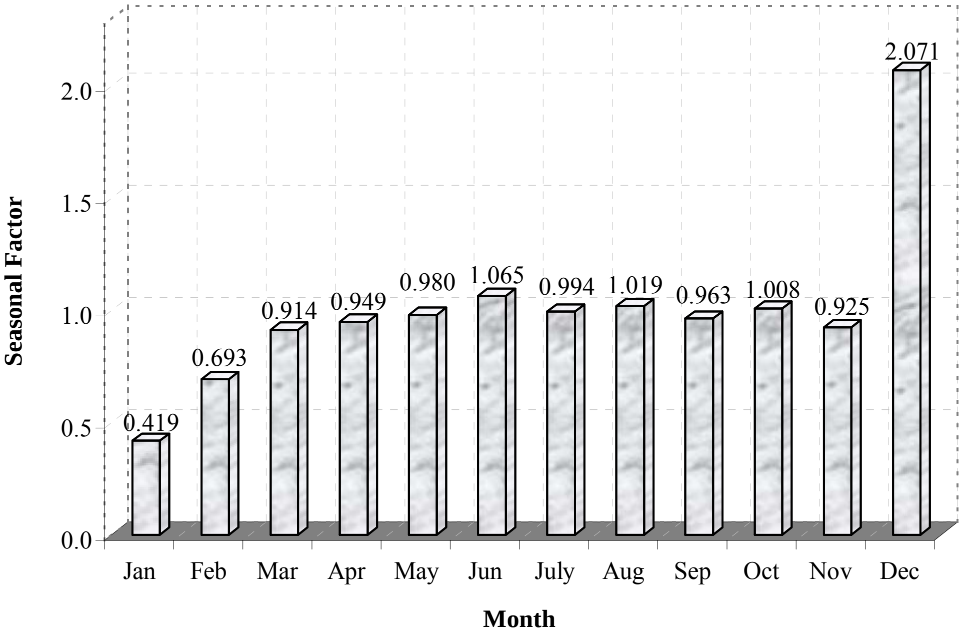


Figure 5.3 Monthly Seasonal factors for US incidence of Tuberculosis

The next step is to calculate the deseasonalized observation in time period *t* as in equation (5.7) and deseasonalized plot is shown in Figure 5.4.

$$d_t = y_t / sn_t \tag{5.7}$$

It is necessary to calculate the seasonality first because it is difficult to identify the trend of the time series. Using the seasonal factors, data can be deseasonalized using equation (5.7) by dividing one monthly data with its respective seasonal factor. Figure 5.4 presents the deseasonalized chart of the time series. This figure compares the original observation (grey line) and the deseasonalized data (black line). From the figure, the deseasonalized data may remove the seasonal factor for the time series. Once, the deseasonalized data have been calculated, the trend of the time series can be determined. The estimation of tr_t for the trend TR_t could be obtained by fitting a regression equation to the deseasonalized data. It yielded function of tr_t.

$$tr_t = b_0 + b_1 t = 2021.734 - 6.012t \tag{5.8}$$

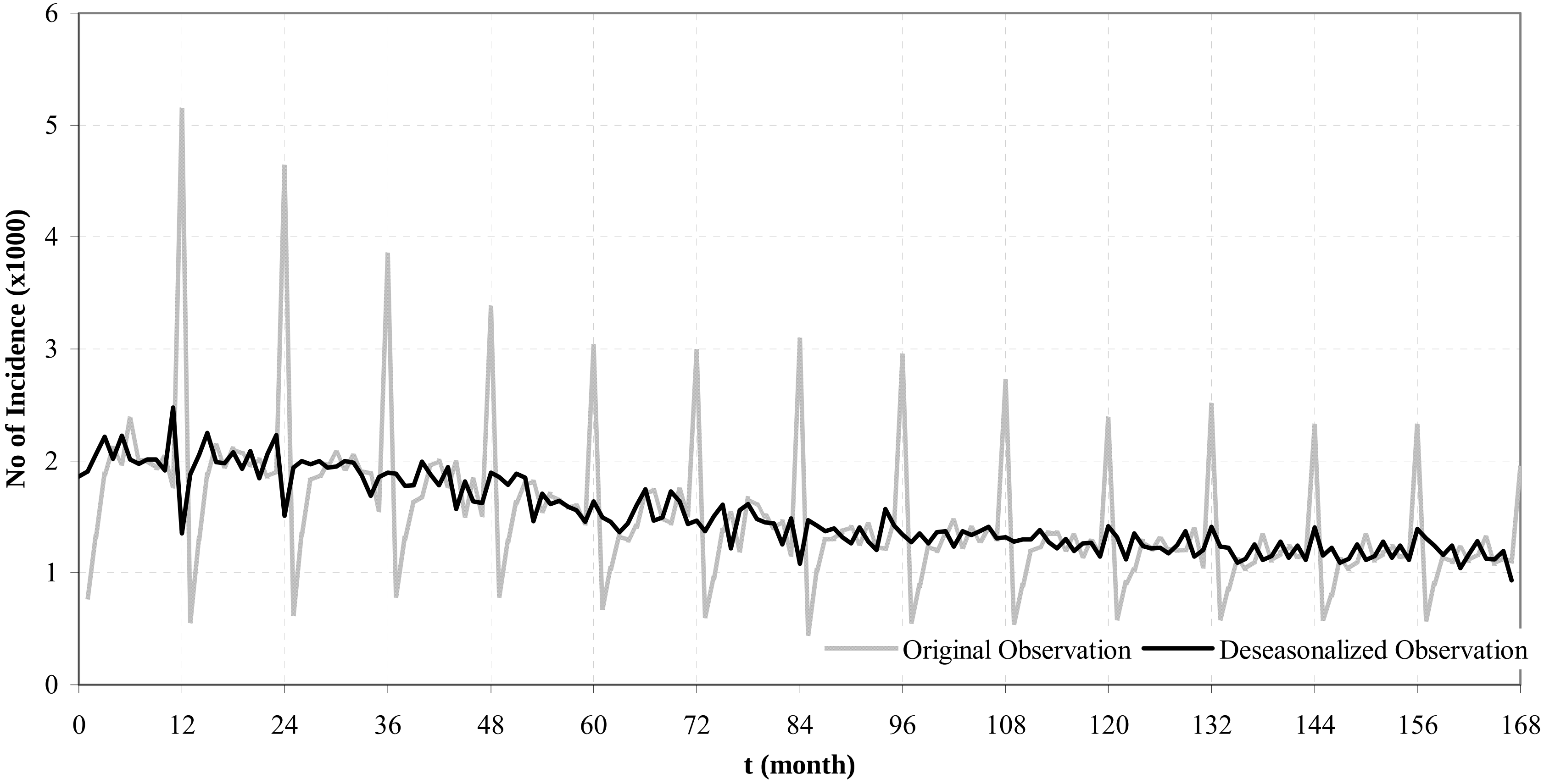


Figure 5.4 Deseasonalized Plot of Tuberculosis

By applying equation (5.8) into each point of time, trend of that month could be calculated. Finally, the final forecasted results could be achieved using equation (5.3). However the irregular component could not be calculated, thus this component had to be removed from the calculation. The final result was obtained by adding seasonality values and trend for the month. The results of the forecasted values in 2008 and 2009 are presented in Section 5.2.7 and the chart of the forecasting results can be found in Appendix B.

5.2.4 Holt-Winter's Result

A Holt-Winter's model required three main smoothing parameters: alpha (α) for level, beta (β) for trend and gamma (γ) for seasonality. Due to the decreasing trend of the time series, the multiplicative Holt-Winter's was selected. Method chosen for determining α, β, and γ was based on mean absolute percentage error (MAPE).

To obtain the fitted parameters that give the best output, several trial and errors were done by using iteration which the starting parameter was 0.01 and incremented by 0.01. Monthly numbers of incidence were used for the forecast model based on formula (3.5) to (3.8). Hence, *s* was chosen to be 12 (number of months in a year).

Table 5.4 Ten Tuberculosis Model with the Smallest Mean Squares

Model	**Alpha**	**Beta**	**Gamma**	**Mean Abs % Error (MAPE)**
451	0.05	0.01	0.01	6.563
676	0.07	0.01	0.01	6.468
901	0.09	0.01	0.01	6.434
466	0.05	0.03	0.01	6.604
452	0.05	0.01	0.03	6.604
691	0.07	0.03	0.01	6.527
227	0.03	0.01	0.03	6.870
1126	**0.11**	**0.01**	**0.01**	**6.422**
916	0.09	0.03	0.01	6.500
228	0.03	0.01	0.05	6.811

Table 5.4 displays parameters for ten smallest mean squares among the iterations. Every model presents the value of alpha (α), beta (β), gamma and mean absolute percentage error (MAPE). The best model is achieved by model 1126 which have the lowest MAPE among the model (MAPE = 6.422). The selected parameter was determined by the smallest MAPE with the result: $\alpha = 0.11$, $\beta = 0.01$, and $\gamma = 0.01$.

Once the best values of alpha, beta, and gamma values were chosen, the forecast model was built using fourteen years of monthly data. Using the model, a monthly incidence forecast was calculated for 2008 and 2009 as shown in Section 5.2.7 and the actual versus forecasted monthly incidence is found in Appendix B.

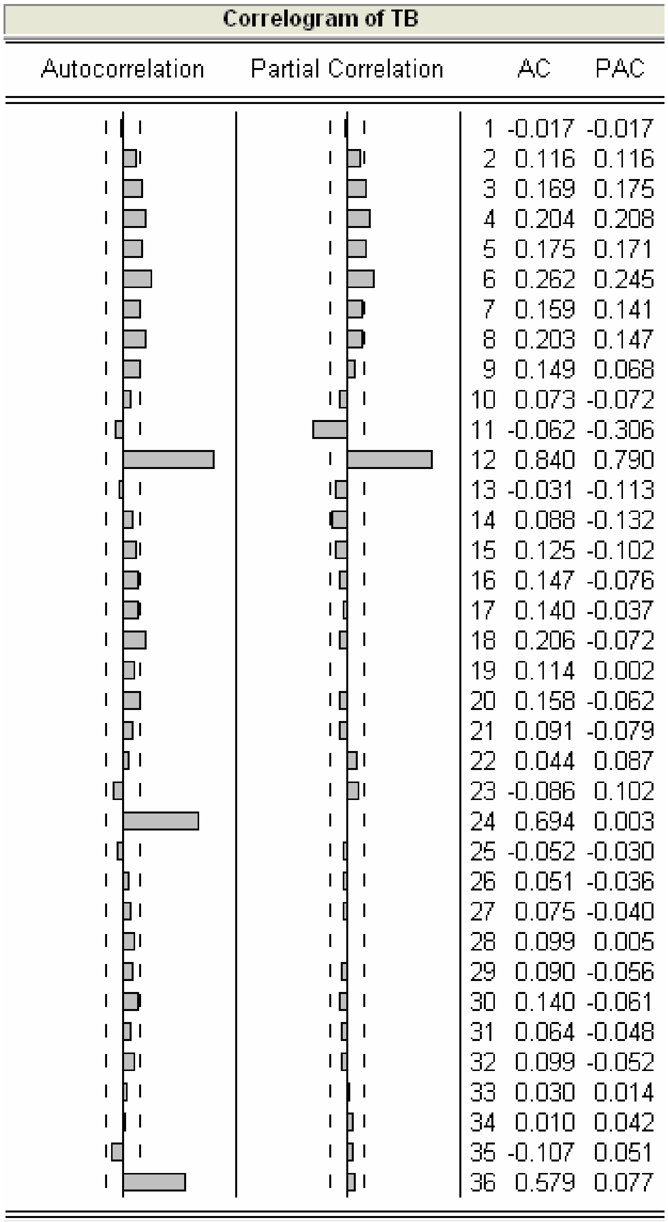
5.2.5 ARIMA Result

This section presents ARIMA results based on BJ steps to predict the future number of Tuberculosis incidence based on the available time series.

5.2.5.1 Identification

In the first step, SAC and SPAC of historical data were plotted to observe the pattern. Three periodical data were selected to illustrate the plot. The correlogram result of original data is shown in Figure 5.5. Based on Figure 5.5, it could be observed that the correlogram of time series is likely to have seasonal cycles especially in SAC which implied level non-stationary.

According to Figure 5.5, it is shown that the time series correlogram likely to have seasonal cycles and shows non-stationarity since the series does not appear vertical along the Y-axis, especially in the SAC correlogram. The non-stationarity of time series can be identified from the values of SAC and SPAC, if the values > |0.2| then the time series is likely not stationarity. SAC values exhibit spikes with the values > |0.2| at its seasonal and near seasonal lag (lag 11, 12, 13, 23, 24, 25, 35, and 36). In the SPAC correlogram, spikes are identified most in the first period (lag 1, 2, 3, 4, 5, 10, 11, 12, and 13). The Augmented Dickey-Fuller (ADF) unit root test was conducted to identify the need of time series differencing.

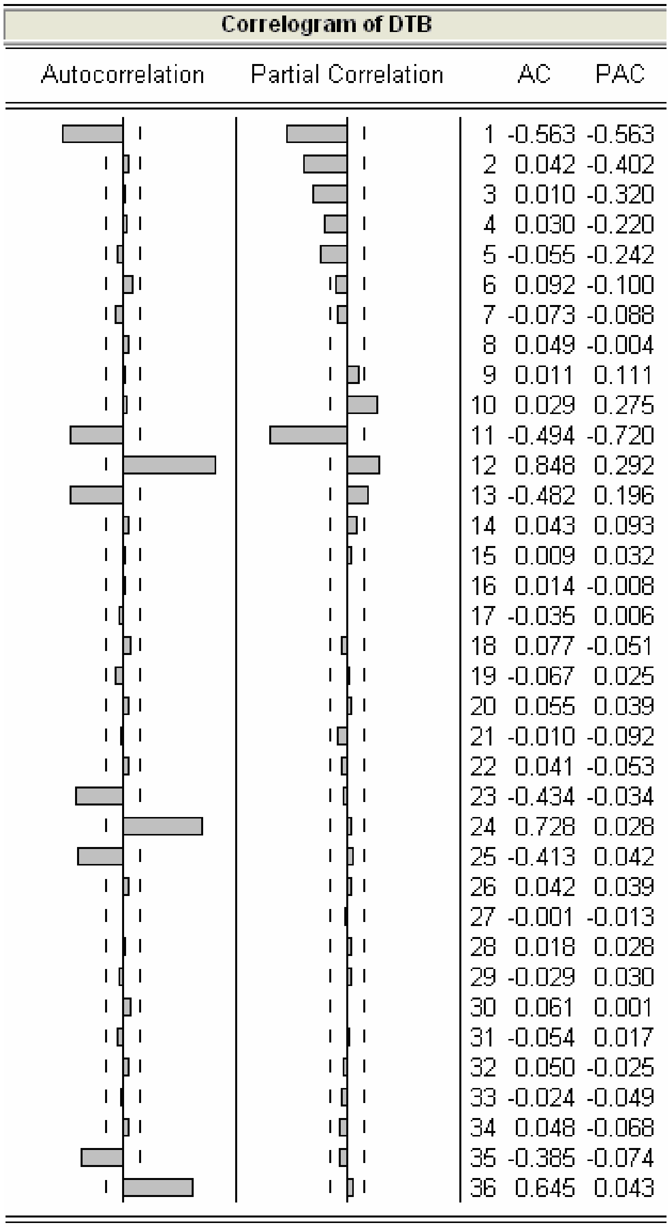


Correlogram of TB

Autocorrelation	Partial Correlation		AC	PAC
		1	-0.017	-0.017
		2	0.116	0.116
		3	0.169	0.175
		4	0.204	0.208
		5	0.175	0.171
		6	0.262	0.245
		7	0.159	0.141
		8	0.203	0.147
		9	0.149	0.068
		10	0.073	-0.072
		11	-0.062	-0.306
		12	0.840	0.790
		13	-0.031	-0.113
		14	0.088	-0.132
		15	0.125	-0.102
		16	0.147	-0.076
		17	0.140	-0.037
		18	0.206	-0.072
		19	0.114	0.002
		20	0.158	-0.062
		21	0.091	-0.079
		22	0.044	0.087
		23	-0.086	0.102
		24	0.694	0.003
		25	-0.052	-0.030
		26	0.051	-0.036
		27	0.075	-0.040
		28	0.099	0.005
		29	0.090	-0.056
		30	0.140	-0.061
		31	0.064	-0.048
		32	0.099	-0.052
		33	0.030	0.014
		34	0.010	0.042
		35	-0.107	0.051
		36	0.579	0.077

Figure 5.5 SAC and SPAC Correlogram of Tuberculosis Original Data

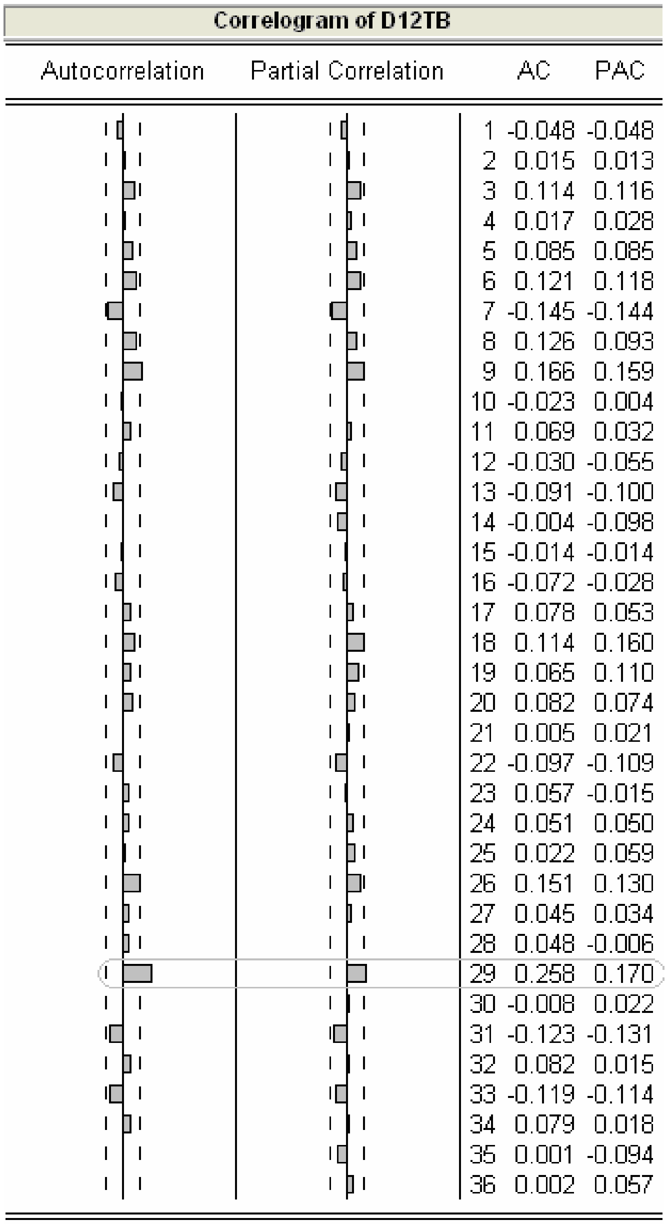
The Augmented Dickey-Fuller (ADF) unit root test was conducted to determine the need of data differencing. The Augmented Dickey-Fuller (ADF) test resulted -3.068 and the one-sided *p*-value is 0.031. The critical values reported at 1%, 5%, and 10% were -3.476, -2.880, -2.577. It showed that t_α value was greater than the critical values that provide evidence not to reject the null hypothesis of a unit root where the data need to be differenced. Therefore, the regular differencing and seasonal differencing was applied to the original time series.



Correlogram of DTB

Autocorrelation	Partial Correlation		AC	PAC
		1	-0.563	-0.563
		2	0.042	-0.402
		3	0.010	-0.320
		4	0.030	-0.220
		5	-0.055	-0.242
		6	0.092	-0.100
		7	-0.073	-0.088
		8	0.049	-0.004
		9	0.011	0.111
		10	0.029	0.275
		11	-0.494	-0.720
		12	0.848	0.292
		13	-0.482	0.196
		14	0.043	0.093
		15	0.009	0.032
		16	0.014	-0.008
		17	-0.035	0.006
		18	0.077	-0.051
		19	-0.067	0.025
		20	0.055	0.039
		21	-0.010	-0.092
		22	0.041	-0.053
		23	-0.434	-0.034
		24	0.728	0.028
		25	-0.413	0.042
		26	0.042	0.039
		27	-0.001	-0.013
		28	0.018	0.028
		29	-0.029	0.030
		30	0.061	0.001
		31	-0.054	0.017
		32	0.050	-0.025
		33	-0.024	-0.049
		34	0.048	-0.068
		35	-0.385	-0.074
		36	0.645	0.043

Figure 5.6 SAC and SPAC Correlogram of Tuberculosis Regular Differencing



Correlogram of D12TB

Autocorrelation	Partial Correlation		AC	PAC
		1	-0.048	-0.048
		2	0.015	0.013
		3	0.114	0.116
		4	0.017	0.028
		5	0.085	0.085
		6	0.121	0.118
		7	-0.145	-0.144
		8	0.126	0.093
		9	0.166	0.159
		10	-0.023	0.004
		11	0.069	0.032
		12	-0.030	-0.055
		13	-0.091	-0.100
		14	-0.004	-0.098
		15	-0.014	-0.014
		16	-0.072	-0.028
		17	0.078	0.053
		18	0.114	0.160
		19	0.065	0.110
		20	0.082	0.074
		21	0.005	0.021
		22	-0.097	-0.109
		23	0.057	-0.015
		24	0.051	0.050
		25	0.022	0.059
		26	0.151	0.130
		27	0.045	0.034
		28	0.048	-0.006
		29	0.258	0.170
		30	-0.008	0.022
		31	-0.123	-0.131
		32	0.082	0.015
		33	-0.119	-0.114
		34	0.079	0.018
		35	0.001	-0.094
		36	0.002	0.057

Figure 5.7 SAC and SPAC Correlogram of Tuberculosis Seasonal Differencing

Figures 5.6 and Figure 5.7 shows the correlogram of Tuberculosis regular time series differencing, and seasonal time series differencing, respectively. In these figures, autocorrelation and partial correlation plot for each time lag is presented next to each unit in column SAC and SPAC. The values of SAC and SPAC that greater than |0.2| indicates a spike. The selection of whether to use regular or seasonal differencing was based on the correlogram. Based on the regular correlograms in Figure 5.6, for SAC there are still some spikes at the first lag, at exact seasonal lags (lag 12, 24, and 36) and at near seasonal lags (lag 11, 13, 23, 25, and 35). Whilst for SPAC of regular differencing, all spikes are found in the first period (lag 1, 2, 3, 4, 5, 10, 11, and 12). For seasonal differencing of Tuberculosis data in Figure 5.7, the only spike is placed at lag 29 of SAC and no spikes are found in the SPAC. Therefore, a seasonal differencing of Tuberculosis data is selected for model development.

5.2.5.2 Parameter Estimation

Different ARIMA models were applied to find the best fitting model. The most appropriate model was selected by using the BIC and AIC values. These models were selected using Maximum Likelihood principle to choose highest possible dimension. The best model was determined from the minimum BIC and AIC. Table 5.5 presents the results of estimating the various ARIMA processes for the seasonal differencing of Tuberculosis human incidence using the EViews 5.1 econometric software package.

Table 5.5 Estimation of Selected SARIMA Model

No	**Model Variable**	**BIC**	**AIC**	**Adj. R^2**
1	C, MA(29)	12.499	12.461	0.412
2	C, AR(29)	12.799	12.754	0.099
3	C, AR(7), SAR(12), MA(7), SMA(12)	12.588	12.482	0.420
4	C, AR(7), AR(29), SAR(12), MA(7), MA(29), SMA(12)	12.055	11.888	0.627
5	C, AR(7), AR(29), SAR(12), MA(29), SMA(12)	12.051	11.908	0.605
6	C, AR(29), SAR(12), MA(29) SMA(12)	12.049	11.930	0.605

The AIC and BIC are commonly used in model selection, whereby the smaller value is preferred [160]. From Table 5.5, model 4 has a relatively small value of BIC and AIC. It also achieved large adjusted R^2. While model 6 has the smallest BIC, but the AIC and adjusted R^2 for this model were larger than model 4. Thus model 4 was selected as the fittest model.

The selected model C AR(7) AR(29) SAR(12) MA(7) MA(29) SMA(12) also can be written as ARIMA(29,0,29)(12,1,12)$_{12.}$ The model yields two AR results (AR(7) and AR(29)) and two MA results (MA(7) and MA(29)) in the regular component. Herein after, the highest lag is selected in the notation and it become ARIMA(29,0,29). For the seasonal component, ARIMA(12,1,12)$_{12}$ is presented because the model consists of SAR(12) and SMA(12). The subscript number '12' shows that there is twelve lags in one period. According to the seasonal differencing used in model, a differencing level of the regular component is 0 and on the contrary a differencing level of the seasonal component is 1.

To produce the model, the separated non-seasonal and seasonal models were computed first. It was followed by combining these models to describe the final model.

- Step 1: Model for nonseasonal level

AR(7) : $z_t = \delta + \phi_7 z_{t-7} + a_t$ (5.9)

AR(29) : $z_t = \delta + \phi_{29} z_{t-29} + a_t$ (5.10)

MA(7) : $z_t = \delta + a_t - \theta_7 a_{t-7}$ (5.11)

MA(29) : $z_t = \delta + a_t - \theta_{29} a_{t-29}$ (5.12)

- Step 2: Model for seasonal level

AR (12) : $z_t = \delta + \phi_{1,12} z_{t-12} + a_t$ (5.13)

MA (12) : $z_t = \delta + a_t - \theta_{1,12} a_{t-12}$ (5.14)

- Step 3: Combining the two present the overall model

$$z_t = \delta + \phi_7 z_{t-7} + \phi_{29} z_{t-29} - \theta_7 a_{t-7} - \theta_{29} a_{t-29} + \phi_{1,12} z_{t-12} - \theta_{1,12} a_{t-12} + a_t \tag{5.15}$$

Hence, δ value is C from the parameter. When an autoregressive and a moving average model are presented in the non-seasonal or seasonal level, multiplicative terms are used as the following:

- $\phi_7 z_{t-7}$, $\phi_{29} z_{t-29}$ and $\phi_{1,12} z_{t-12}$ were used to form the multiplicative term $\phi_7 \phi_{1,12} z_{t-19}$ and $\phi_{29} \phi_{1,12} z_{t-41}$
- $-\theta_7 a_{t-7}$, $-\theta_{29} a_{t-29}$ and $-\theta_{1,12} a_{t-12}$ were used to form the multiplicative term $\theta_7 \theta_{1,12} a_{t-19}$ and $\theta_{29} \theta_{1,12} a_{t-41}$

The model was derived using multiplicative form as described in (5.16):

$$\begin{aligned} z_t &= \delta + \phi_7 z_{t-7} + \phi_{29} z_{t-29} - \theta_7 a_{t-7} - \theta_{29} a_{t-29} + \phi_{1,12} z_{t-12} - \theta_{1,12} a_{t-12} + \phi_7 \phi_{1,12} z_{t-19} \\ &\quad + \phi_{29} \phi_{1,12} z_{t-41} + \theta_7 \theta_{1,12} a_{t-19} + \theta_{29} \theta_{1,12} a_{t-41} + a_t \\ z_t &- \phi_7 z_{t-7} - \phi_{29} z_{t-29} - \phi_{1,12} z_{t-12} - \phi_7 \phi_{1,12} z_{t-19} - \phi_{29} \phi_{1,12} z_{t-41} \\ &= \delta + \theta_7 a_{t-7} + \theta_{29} a_{t-29} + \theta_{1,12} a_{t-12} - \theta_7 \theta_{1,12} a_{t-19} - \theta_{29} \theta_{1,12} a_{t-41} + a \end{aligned} \tag{5.16}$$

The final tentatively model by using backshift (B) operator was:

$$\begin{aligned} &(1 - \phi_7 B^7 - \phi_{29} B^{29} - \phi_{1,12} B^{12} - \phi_7 \phi_{1,12} B^{19} - \phi_{29} \phi_{1,12} B^{41}) z_t \\ &\quad = \delta + (1 + \theta_7 B^7 + \theta_{29} B^{29} + \theta_{1,12} B^{12} - \theta_7 \theta_{1,12} B^{19} - \theta_{29} \theta_{1,12} B^{41}) a_t \end{aligned} \tag{5.17}$$

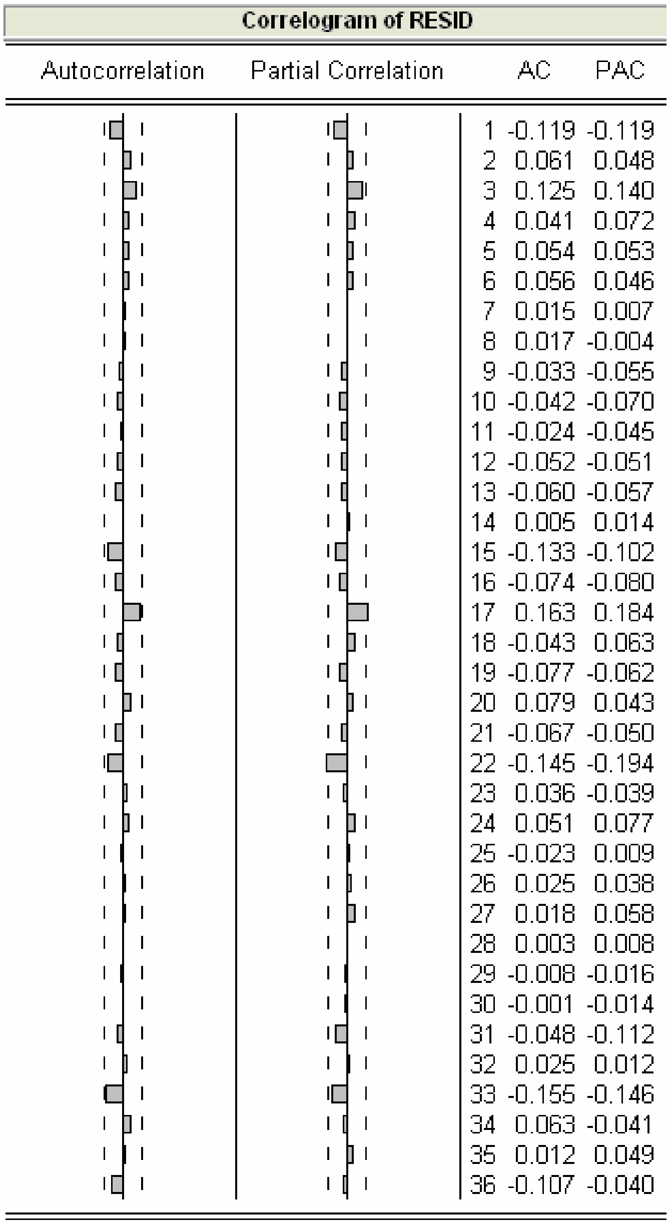
From this model, the parameter results are C = -19.347, AR(7) = 0.197, AR(29) = 0.374, SAR(12) = 0.293, MA(7) = -0.945, MA(29) = 0.0295, SMA(12) = -0.884. The estimated parameters were substituted into equation (5.17) to form the final model that could be expressed as follows:

$$\begin{aligned} &(1 - 0.197B^7 - 0.374B^{29} - 0.293B^{12} - 0.058B^{19} + 0.110B^{41}) z_t \\ &\quad = -19.347 + (1 - 0.945B^7 + 0.029B^{29} - 0.884B^{12} + 0.835B^{19} - 0.026B^{41}) a_t \end{aligned} \tag{5.18}$$

Since the seasonal differencing was chosen, then z_t was defined as (4.17). The SARIMA final model was used to compute the forecast values for a year-ahead.

5.2.5.3 Diagnostic Checking

Diagnostic checking was performed into the selected model. The correlogram and residual plots are presented in Figure 5.8 and LM results are listed in Table 5.6. Table 5.6 shows no correlation up to the order 12 because the t-Statistic is less than |0.2|. The results of Figure 5.8 and Table 5.6 demonstrate that the selected model is suitable.



Correlogram of RESID

Autocorrelation	Partial Correlation		AC	PAC
		1	-0.119	-0.119
		2	0.061	0.048
		3	0.125	0.140
		4	0.041	0.072
		5	0.054	0.053
		6	0.056	0.046
		7	0.015	0.007
		8	0.017	-0.004
		9	-0.033	-0.055
		10	-0.042	-0.070
		11	-0.024	-0.045
		12	-0.052	-0.051
		13	-0.060	-0.057
		14	0.005	0.014
		15	-0.133	-0.102
		16	-0.074	-0.080
		17	0.163	0.184
		18	-0.043	0.063
		19	-0.077	-0.062
		20	0.079	0.043
		21	-0.067	-0.050
		22	-0.145	-0.194
		23	0.036	-0.039
		24	0.051	0.077
		25	-0.023	0.009
		26	0.025	0.038
		27	0.018	0.058
		28	0.003	0.008
		29	-0.008	-0.016
		30	-0.001	-0.014
		31	-0.048	-0.112
		32	0.025	0.012
		33	-0.155	-0.146
		34	0.063	-0.041
		35	0.012	0.049
		36	-0.107	-0.040

Figure 5.8 SAC and SPAC Correlogram of Tuberculosis Model Residual

The correlogram in Figure 5.8 is plotted to identify the randomness. The error is found to be white noise, with the element of both SAC and SPAC falling inside the critical values. There is no serial correlation in the residual because the SAC and SPAC values at all lag are nearly zero and are within the 95% confidence interval. There was no serial correlation in the residual because the SAC and SPAC at all lag is nearly zero and is within the 95% confidence interval.

Table 5.6 LM Test Result for Tuberculosis Residual

Variable	Std. Error	t-Statistic	Prob.
RESID(-1)	-0.125597	-1.181009	0.106347
RESID(-2)	-0.089729	-0.844433	0.106260
RESID(-3)	0.104332	0.959781	0.108704
RESID(-4)	0.030576	0.280472	0.109015
RESID(-5)	-0.034590	-0.312605	0.110649
RESID(-6)	0.028174	0.252612	0.111530
RESID(-7)	-0.153888	-1.099726	0.139933
RESID(-8)	-0.041217	-0.357965	0.115142
RESID(-9)	-0.090214	-0.789727	0.114234
RESID(-10)	-0.048517	-0.422515	0.114830

5.2.5.4 Forecasting

ARIMA(29,0,29)(12,1,12)$_{12}$ has been chosen as the fittest model. Therefore, it was used to forecast the values of the 24 months-ahead from t_{169} through t_{180} that are presented in Section 5.2.7.

5.2.6 Neural Network Result

In this method, 168 monthly data were divided into: 12 data for input, 84 data for training, 36 data for selection, and 36 data for testing. There were 1500 network iteration with 600 network retained. Several MLP networks with different combination of hidden layers and neuron were generated and tested to obtain the best fitted network. The three layer neural network was able to obtain the best network at iteration index 450 that having 12 input nodes, 3 hidden nodes, and 1 output node. The selected network achieved is presented in Table 5.7, while the network is in Figure 5.9.

Table 5.7 Tuberculosis Network Performance

Profile	**MLP s12 1:12-3-1:1**
Train Perf.	0.181
Select Perf.	0.196
Test Perf.	0.299
Train Error	0.028
Select Error	0.031
Test Error	0.032
Training/Members	BP100
Inputs	1
Hidden(1)	3
Hidden(2)	0

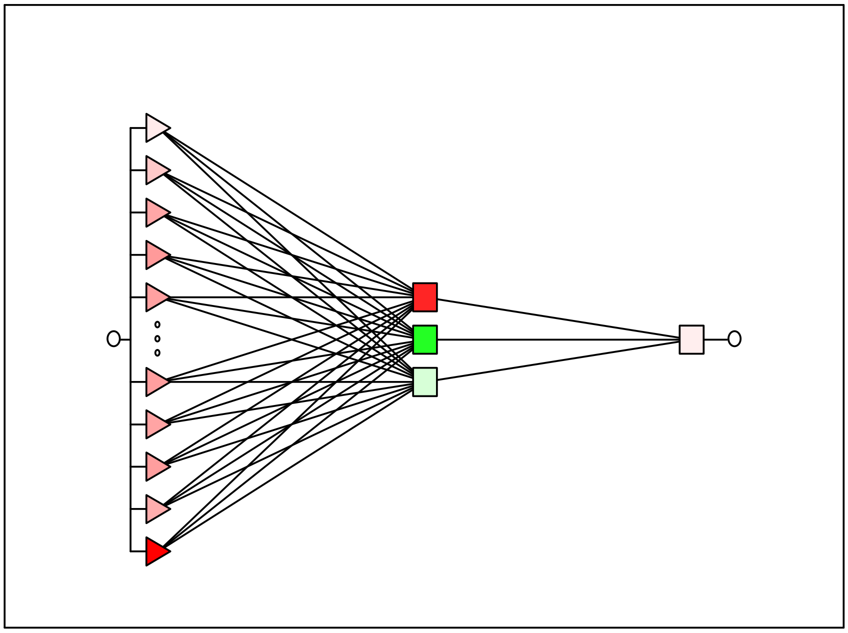


Figure 5.9 Architecture of Tuberculosis Model

Table 5.7 can be described as the following. Profile shows the network type, the number of input and output variables, the number of layers, and the number of neurons in each layer. In Statistica software, the format is <type> <inputs>:<layer1>-<layer2>-<layer3>:<outputs>, where the number of layers may vary. Then, MLP s12 1:12-5-1:1 is defined as a Multilayer Perceptron with one input variable and one output variable, and three layer of input, hidden, output with 12, 3, and 1 unit respectively, whereas it is clearly seen in Figure 5.9. Train perf., select perf., and test perf. show the network performance on the training, selection, and subset test. The values of each performance indicate the ratio of prediction to observation standard deviations. Train error/select error/test error shows error rates of the subsets, while training/members describe the training algorithm used to train the network. According to Table 5.7, BP100 reports one hundred epochs of back propagation.

5.2.7 Forecasting Result

The forecasting results for 2008 and 2009 based on range 1993-2006 for every model are listed in Table 5.8. Figure 5.10 illustrated a sample of completed Tuberculosis time series plots of moving average method. This chart consists of three components: actual values, fitted values, and residual values that are represented by grey solid line, the black dashed line, and the dotted black line, respectively.

The 144 monthly forecasts (January 1994 – December 2006) generated using the moving average model is compared with the actual data. The forecast errors for the entire month are presented as the residual values. The residual values are resulted from the differences between actual data and forecasted value in associated month. Figure 5.10 also provided the predicted number of incidence in 2008 and 2009. The rest of forecasting plots of regression, decomposition, Holt-Winter's, ARIMA, and neural network are given in Figure B2.1, B2.2, B2.3, B2.4, and B2.5 of Appendix B.

Table 5.8 Forecasting Results for Tuberculosis

t	Moving Average	Linear Regression	Decomposition	Holt-Winter’s	ARIMA	Neural Network
169	490.606	52.674	421.165	477.543	639.388	705.817
170	813.664	482.567	693.130	778.735	702.332	1078.202
171	1067.302	828.888	907.802	1027.003	1357.040	997.663
172	1122.131	897.174	937.585	1060.408	1473.149	1099.010
173	1138.614	922.245	962.415	1098.944	931.812	1082.973
174	1223.780	1050.174	1038.645	1181.887	1953.178	1155.825
175	1146.948	932.459	963.929	1096.739	1229.784	1195.335
176	1172.232	970.959	981.961	1122.147	1659.108	1206.835
177	1082.218	868.388	922.247	1060.972	899.529	1098.159
178	1133.939	930.817	959.611	1102.072	1328.327	964.038
179	1024.938	800.317	874.397	1006.739	1383.095	1091.765
180	2282.401	2503.745	1945.765	2229.010	2576.885	1736.163
181	464.290	-23.989	390.954	445.865	615.625	733.548
182	778.601	405.904	643.112	725.867	684.849	1091.572
183	1027.303	752.226	841.896	955.700	1332.413	1000.541
184	1083.067	820.511	869.101	985.133	1453.616	1078.629
185	1094.830	845.583	891.688	1019.213	907.725	1058.650
186	1180.444	973.511	961.845	1094.277	1934.386	1133.220
187	1095.602	855.797	892.212	1013.659	1205.717	1170.706
188	1116.410	894.297	908.447	1035.340	1633.178	1190.171
189	1035.805	791.726	852.770	977.138	881.893	1079.647
190	1083.501	854.154	886.862	1013.149	1303.214	943.550
191	977.494	723.654	807.687	923.821	1362.922	1078.699
192	2187.000	2427.083	1796.367	2041.485	2554.489	1630.433

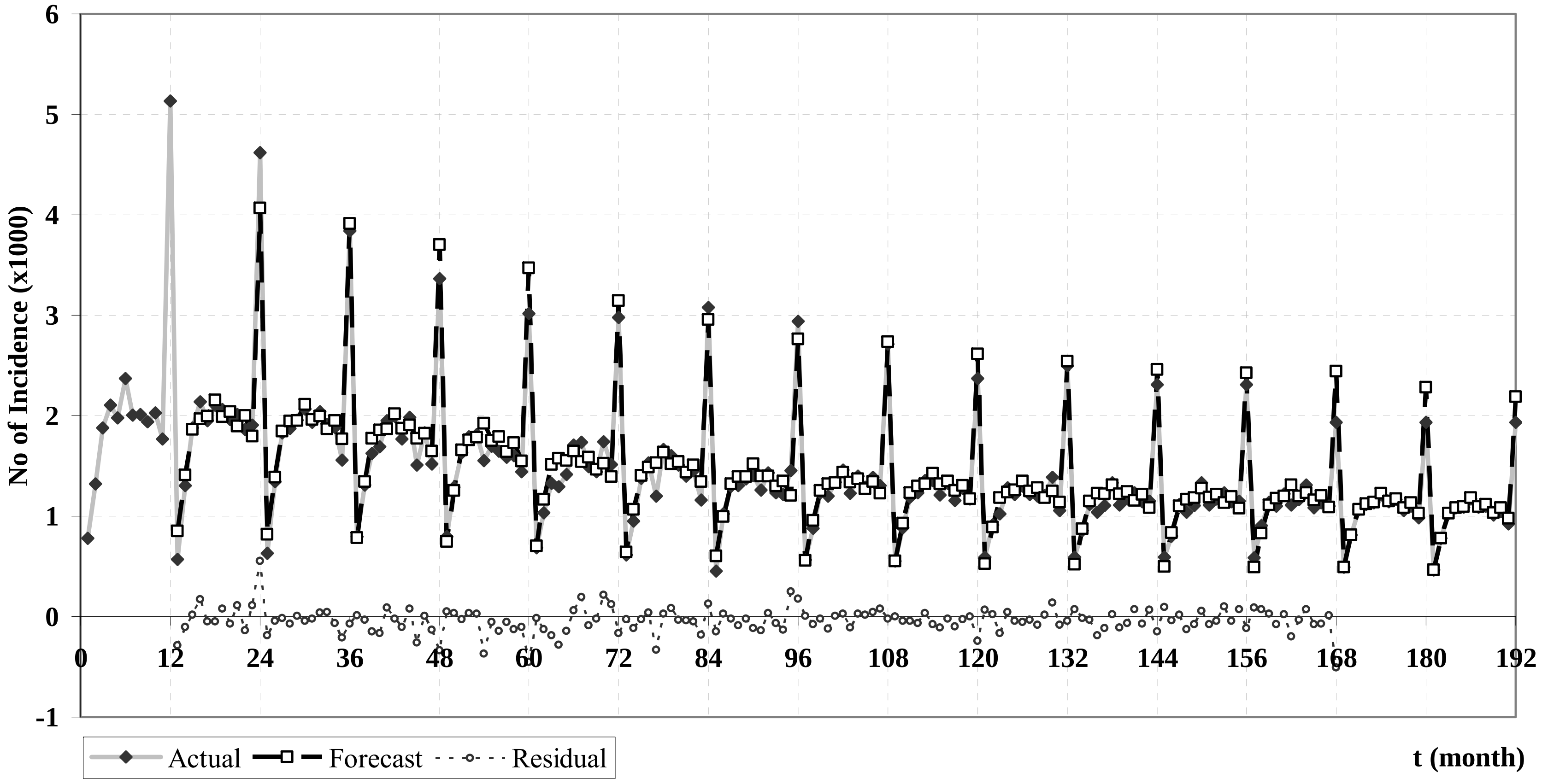


Figure 5.10 Moving Average Result of Tuberculosis

5.2.8 Analysis of Variance (ANOVA)

Six different forecasting methods were applied, namely linear regression, moving average, decomposition, Holt-Winter's, ARIMA, and neural network. The estimated values of them were compared with actual values using ANOVA. The hypothesis was:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$$

$$H_1 : \mu_i \neq \mu_j \quad i, j = 1,2,3,4,5,6,7, \, i \neq j$$

Where $\mu_{1,}$ $\mu_{2,}$ $\mu_{3,}$ $\mu_{4,}$ $\mu_{5,}$ $\mu_{6,}$ μ_7 were the average values obtained from actual data, followed by the average estimation results from moving average, linear regression, decomposition, Holt-Winter's, ARIMA, and neural network, respectively.

The hypotheses was tested by using level of significance (α) assumed as 0.05. The ANOVA results for this problem are shown in Table 5.9. From the ANOVA result shown in Table 5.9, SS (α= 0.05) was 4458661 and MS was 6518.51. It also can be concluded that based on α = 0.05, the null hypothesis was rejected because $F > F_{crit}$, where: F_{crit} = 2.111817 and F = 5.60184.

The P value of between group provide an evidence to reject the null hypothesis, where $P = 0.007344 < 0.05$.

Due to the significant difference between forecasting results, a multiple range test is applied. The next section reports the application of Duncan's Multiple Range Test.

Table 5.9 Blocked Design ANOVA for Tuberculosis

SUMMARY

Groups	**Count**	**Sum**	**Average**	**Variance**
Actual	115	152777.5	1328.5	239490.6
Moving Average	115	157769.2	1371.906	267647
Regression	115	155984.7	1356.388	352614
Decomposition	115	156925.2	1364.567	288544.1
Holt-Winter's	115	153891.7	1338.189	262309.2
ARIMA	115	152436.4	1325.534	252202.4
Neural Network	115	154591.3	1344.272	230763.7

Source of Variation	**SS**	**df**	**MS**	**F**	**P**	**F_{crit} (α=0.05)**
Between Group (Method)	219093.9	6	36515.65	5.60184	1.09E-05	2.111817
Blocks (month)	2.11E+08	114	1854460	284.4914	0	1.251842
Within Groups (Error)	4458661	684	6518.51			
Total	2.16E+08	804				

5.2.9 Duncan Multiple Range Test

This stage tests between forecast results means using Duncan's multiple range tests to check which methods are different.

Let μ_1 represent the mean of the actual data

Let μ_2 represent the mean of moving average result

Let μ_3 represent the mean of regression analysis result

Let μ_4 represent the mean of decomposition result

Let μ_5 represent the mean of holt-winter's result

Let μ_6 represent the mean of ARIMA analysis result

Let μ_7 represent the mean of neural network result

Based on Table 4.9, MS = 6518.51 and n = 115. Standard error of each average,

$$S = \sqrt{\frac{MS}{n}} = 7.529$$

The total number of selected groups was 7, so that the total number of ranges equal to 6. From the table of significant ranges Montgomery for 684 degrees of freedom (684 is the number of degrees of freedom for within groups from ANOVA table) and $\alpha = 0.05$, the six ranges were calculated as given below:

$$r_2 = r_{\alpha(p,df)} = r_{0.05(2,684)} = 2.772$$

$$r_3 = r_{\alpha(p,df)} = r_{0.05(3,684)} = 2.918$$

$$r_4 = r_{\alpha(p,df)} = r_{0.05(3,684)} = 3.017$$

$$r_5 = r_{\alpha(p,df)} = r_{0.05(5,684)} = 3.089$$

$$r_6 = r_{\alpha(p,df)} = r_{0.05(6,684)} = 3.146$$

$$r_7 = r_{\alpha(p,df)} = r_{0.05(7,684)} = 3.193$$

LSR results were:

$$R_2 = r_2 \times S = r_{0.05(2,684)} \times S = 2.772 \times 7.529 = 20.870$$

$$R_3 = r_3 \times S = r_{0.05(3,684)} \times S = 2.918 \times 7.529 = 21.969$$

$$R_4 = r_4 \times S = r_{0.05(4,684)} \times S = 3.017 \times 7.529 = 22.714$$

$$R_5 = r_5 \times S = r_{0.05(5,684)} \times S = 3.089 \times 7.529 = 23.256$$

$$R_6 = r_6 \times S = r_{0.05(6,684)} \times S = 3.146 \times 7.529 = 23.686$$

$$R_7 = r_7 \times S = r_{0.05(7,684)} \times S = 3.193 \times 7.529 = 24.039$$

Every method and its mean value were symbolized in Table 5.10 below.

Table 5.10 Mean between different pair

Methods	**Mean Symbol**	**Mean Value**
Actual	M1	1328.500
Moving Average	M2	1371.906
Regression	M3	1356.388
Decomposition	M4	1364.567
Holt-Winter's	M5	1338.189
ARIMA	M6	1325.534
Neural Network	M7	1347.324

Means values in Table 5.10 was sorted as displayed in Table 5.11

Table 5.11 The Sorted Mean Values of Tuberculosis

Method	**M6**	**M1**	**M5**	**M7**	**M3**	**M4**	**M2**
Mean	1325.534	1328.500	1338.189	1344.272	1356.388	1364.567	1371.906

The differences between different pairs of means were:

M1 – M2 = 43.406 > 20.870 (R_2)

M1 – M3 = 27.888 > 21.969 (R_3)

M1 – M4 = 36.067 > 22.714 (R_4)

M1 – M5 = 9.689 < 23.256 (R_5)

M1 – M6 = 2.966 < 23.686 (R_6)

M1 – M7 = 15.772 < 24.039 (R_7)

M2 – M3 = 15.518 < 20.870 (R_2)

M2 – M4 = 7.339 < 21.969 (R_3)

M2 – M5 = 33.717 > 22.714 (R_4)

M2	–	M6	=	46.372	>	23.256	(R_5)
M2	–	M7	=	27.634	>	23.686	(R_6)
M3	–	M4	=	8.179	<	20.870	(R_2)
M3	–	M5	=	18.200	<	21.969	(R_3)
M3	–	M6	=	30.855	>	22.714	(R_4)
M3	–	M7	=	12.116	<	23.256	(R_5)
M4	–	M5	=	26.378	>	20.870	(R_2)
M4	–	M6	=	39.033	>	21.969	(R_3)
M4	–	M7	=	20.295	<	22.714	(R_4)
M5	–	M6	=	12.655	<	20.870	(R_2)
M5	–	M7	=	6.083	<	21.969	(R_3)
M6	–	M7	=	18.738	<	20.870	(R_2)

The actual differences between the different pairs of means could be converted into Figure 5.11. Figure 5.11 showed the actual difference (AD) between forecast means. The sorted means was presented at the top of figure that started from M6, where M1 = mean of the actual data, M2 = mean of moving average result, M3 = mean of regression forecasting result, M4 = mean of decomposition method, M5 = mean of the Holt-Winter's method, M6 = mean of ARIMA result, and M7 = mean of Neural Network. ARIMA had the smallest mean among them, and then it located at the left side. This is followed by the actual data mean, the Holt-Winter's mean, the Neural Network mean, the regression mean, the decomposition mean, while the regression mean which had the largest mean was located at the right side.

The AD for the associated pair, which has a significant difference from the calculation, is presented by using bold font and grey background as shown in Figure 5.10. The first six AD show the differences between actual data and the forecasting result from different methods. Three forecasting methods (regression, decomposition, and moving average) resulted in differences with actual data. AD of these three methods are greater than their least significant range, where M1 – M2 = 43.406 > 20.870 for moving average, M1 – M3 = 27.888 > 21.969 for regression, and M1 – M4 = 36.067 > 22.714 for decomposition.

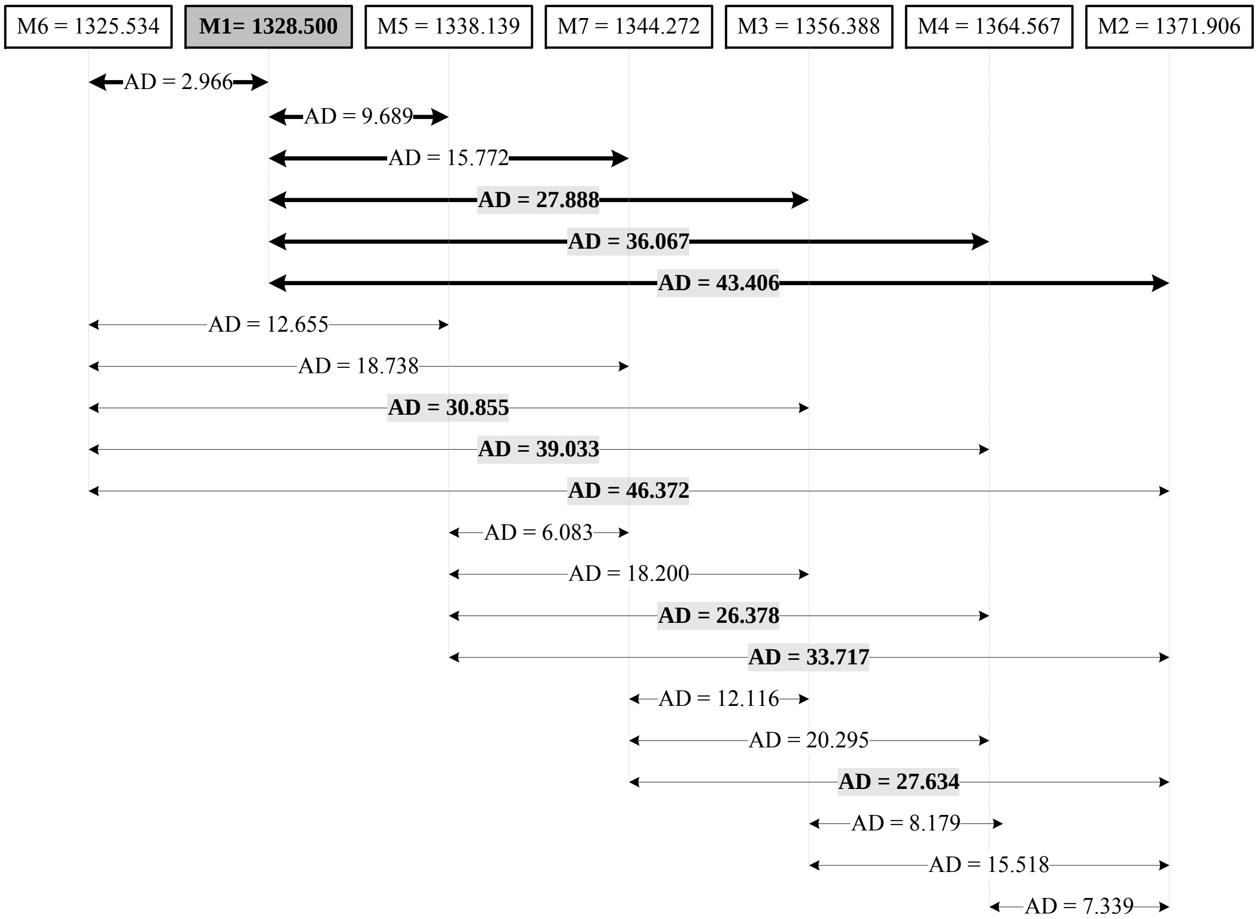


Figure 5.11 Actual Difference between Different Pairs of Means while Comparing Different Forecasting Method

The moving average, regression, and decomposition methods also perform badly compared to the other techniques. The mean difference of moving average, regression, decomposition also have differences with others (Holt-Winter's, ARIMA, neural network) because their AD are greater than the least significant range (M2 – M4 = 106.738 > 72.437, M3 – M4 = 107.843 > 68.812, M4 – M5 = 100.508). This indicates that moving average, regression, and decomposition methods are not appropriate to be applied to the Tuberculosis data.

In contrast, no difference was observed from the mean comparison between the actual data and the rest of the methods (Holt-Winter's, ARIMA, and neural network). Since there were no significant difference between actual data and these methods, it is concluded that the Holt-Winter's, ARIMA, and neural network could be used mainly for predicting future number of human incidence using the available Tuberculosis data.

5.2.10 Performance of Forecast Results

Holt-Winter's, ARIMA, and neural network were identified as the suitable methods to Tuberculosis time series. Therefore, coefficient of variation (CV) was used to rank the methods and determined the fittest method among them. The outputs of CV values are shown in Table 5.12.

Table 5.12 CV Values of Forecast Results

Method	**Standard Deviation (σ)**	**Mean (μ)**	**Coefficient of Variation (CV)**	**Rank**
Holt-Winter's	512.1613	1338.189	0.383	3
ARIMA	502.1976	1325.534	0.379	2
Neural Network	480.3787	1344.272	0.357	1

Table 5.12 presents the standard deviation, the mean, the coefficient of variation (CV) and rank of each method. The method ranked table is based on the value of CV. A high CV value reflects inconsistency among the samples within the group of forecast results. Thus, method with the smallest CV is identified as the top ranking.

Referring to Table 4.11, highest rank is achieved by neural network, followed by ARIMA, and the lowest rank is Holt-Winter's.

5.3 Dialog Generation and Management Subsystem

'What if analysis' was conducted to determine the fluctuation in the forecasting results in a specific method that is influenced by changes in the historical Tuberculosis data. To show this analysis, outline of steps were organized as described in Section 4.3.

Table 5.13 Sensitivity of Moving Average Tuberculosis

Month	Forecast 2008			Forecast 2009		
	Based data 1993-2006	Based data 1993-2007	Sensitivity (%)	Based data 1993-2006	Based data 1993-2007	Sensitivity (%)
January	464.290	480.053	3.28%	444.883	467.178	4.77%
February	778.266	774.471	-0.49%	746.625	772.278	3.32%
March	1026.103	1021.657	-0.44%	984.545	1019.730	3.45%
April	1080.693	1081.516	0.08%	1036.279	1074.848	3.59%
May	1090.927	1117.981	2.42%	1044.771	1096.600	4.73%
June	1173.505	1200.846	2.28%	1122.969	1179.122	4.76%
July	1086.138	1124.103	3.38%	1037.279	1095.082	5.28%
August	1099.786	1154.542	4.74%	1051.669	1115.379	5.71%
September	1009.842	1083.709	6.82%	969.448	1035.172	6.35%
October	1044.209	1108.611	5.81%	1003.706	1065.108	5.76%
November	921.145	1009.240	8.73%	891.003	950.860	6.30%
December	1936.000	1983.000	2.37%	1936.000	1983.000	2.37%

Table 5.13 summarizes sensitivity analysis for moving average results. The table is separated into two groups. The first group presents the percentage of sensitivity analysis in forecasting number of incidence in 2008. The percentage was obtained from the difference between forecast result between 1993–2006 data range and 1993–2007 data range as mentioned in Equation (3.33). Furthermore, the second group shows the sensitivity analysis of the forecasted number of incidence in 2009 that were

resulted from the same formula in the preceding column. The table of sensitivity analysis for other methods can be seen in the sensitivity analysis of other methods is given in Appendix C, where Table C2.1 for regression, Table C2.2 for decomposition, Table C2.3 for Holt-Winter's, Table C2.4 for ARIMA and Table C2.5 for neural network.Appendix C.

The positive sensitivity indicates that forecast values in range 1993–2007 is higher than values resulted from range seems 1993– 2006. While, the negative sensitivity exhibits that the result in range 1993– 2007 is lower than in range 1993–2007. Figure 5.12 plots the monthly forecast for 2008 based on two dataset; one based on the data from 1993 – 2006 and the other based on the data from 1993 – 2007. Whilst, Figure 5.13 plots the monthly forecast for 2009 based on two dataset. The sensitivity plots of the regression, decomposition, Holt-Winter's, ARIMA, and neural network can be found in Table D2.1, Table D2.2, Table D2.3, Table D2.4, and Table D2.5 of Appendix D.

5.4 Summary

Chapter 5 reports an overall result of the development of DSS framework in multiplicative seasonal model. To achieve this goal, Tuberculosis time series data was used. US data of Tuberculosis incidence indicated a decreasing trends over years, hence the reason of Tuberculosis being chosen. As in Salmonellosis, three components of DSS framework was developed by using Tuberculosis data set. Although some part utilizes different approaches from Salmonellosis study, the final result could be directed and be used with the proposed framework especially using multiplicative time series data. The results of Chapter 4 and Chapter 5 are compared in Chapter 6 to identify the similarities and the differences between them.

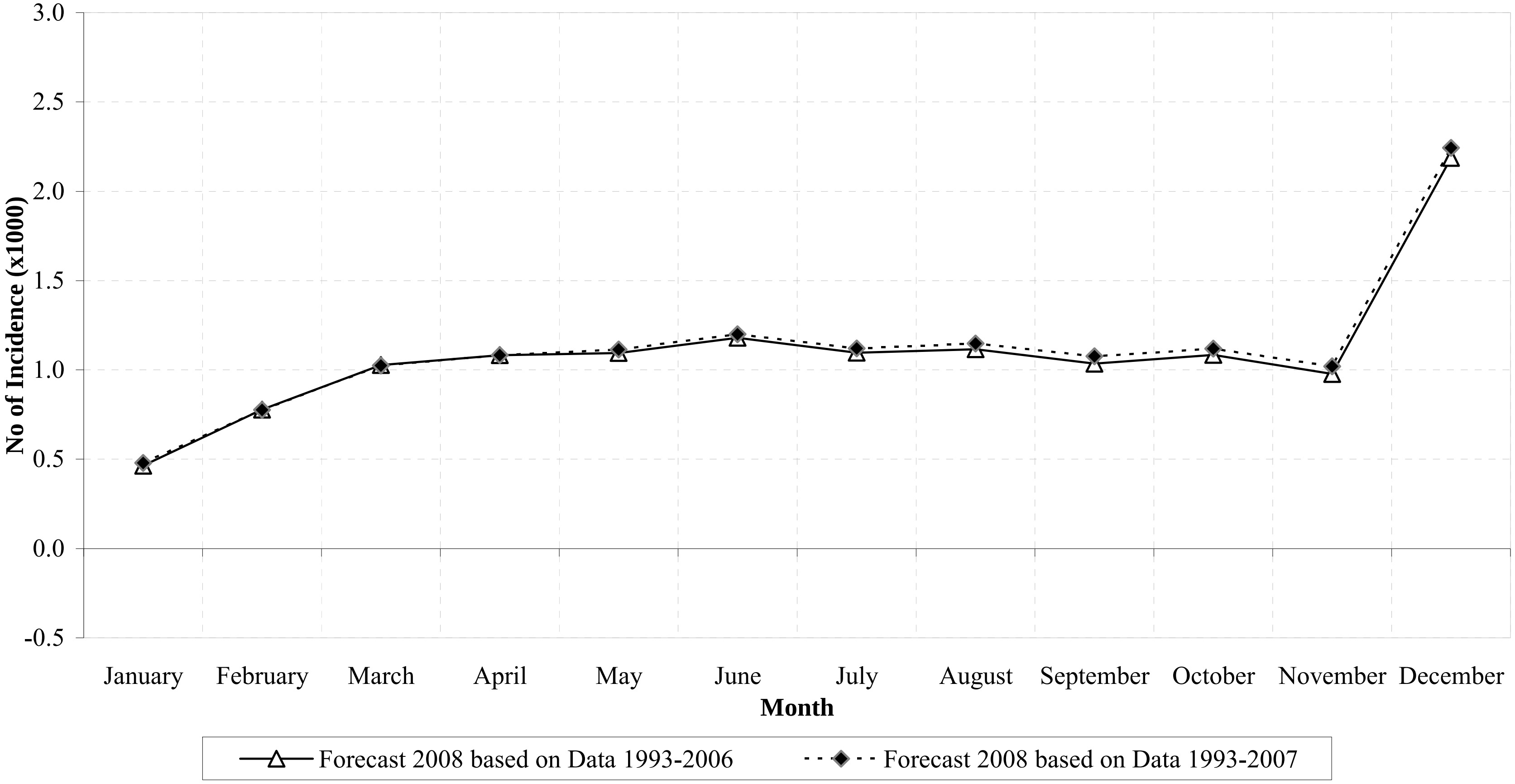


Figure 5.12 Sensitivity Analysis of Moving Average Forecast 2008 for Tuberculosis

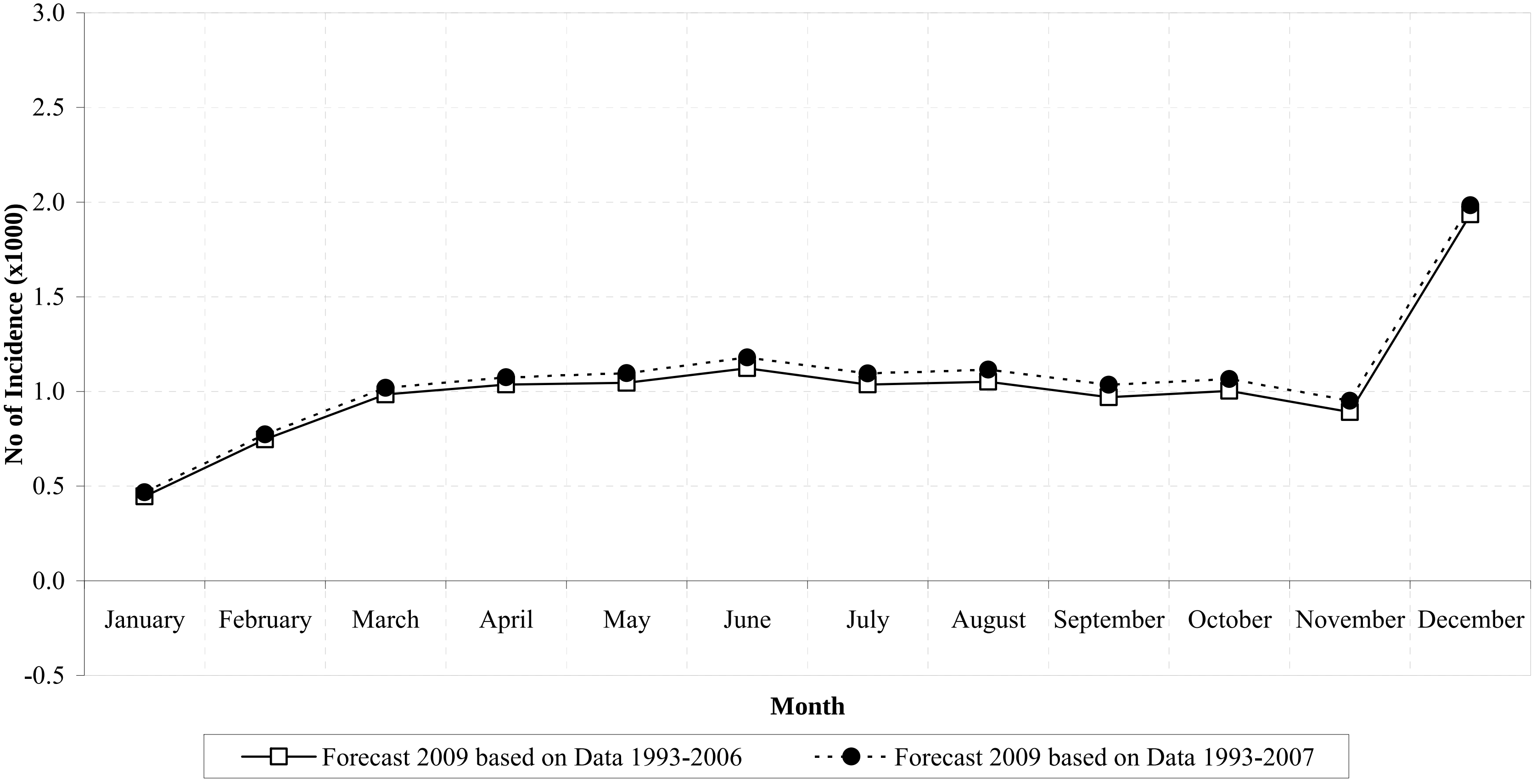


Figure 5.13 Sensitivity Analysis of Moving Average Forecast 2009 for Tuberculosis

CHAPTER 6
DISCUSSION

6.0 Chapter Overview

The iterative process described in Chapter 3 (research methodology) was conducted for each case study throughout the searching of various literatures and based on the problem statement. The problem statement was identified and a general framework that was focused in conceptual and mathematical formula of a related model was developed. The guidelines directed the research development and provided an understanding of model steps.

The DSS framework composed of three components. Once the general framework was developed, two case studies were applied on the framework, as presented separately in Chapter 4 and Chapter 5. A comparative analysis was conducted between the findings of case 1 and case 2. This chapter looks across the results from the two case studies to compare and to highlight the similarities, differences, and specific patterns found. It also provides the analysis for identifying the causal factors for differences between the case studies.

The discussion of DSS components with respect to both cases are presented in Section 6.1, Section 6.2, and Section 6.3 for database management subsystem, model base management subsystem, and dialog generation management subsystem, respectively. These are followed by a discussion of the cross case analysis in Section 6.4, a design of Graphical User Interface (GUI) in Section 6.5 and a presentation of framework benefit in Section 6.6. The final section of this chapter, Section 6.7 provides the chapter summary.

6.1 Analysis of Database Management Subsystem

The DSS models represented the application of the proposed framework in two different case study models. Case study 1 consisted of a zoonosis DSS model that was applied based on the additive time series. Case study 2 consisted of a model that was structured based on the multiplicative time series. In general, it is seen that the DSS is able to be applied to both time series model, as reported in Chapter 4 and Chapter 5.

The time series of both database management subsystems were collected from the same sources as listed in Section 3.5.3. The fifteen annual raw data were collected and divided into monthly data, to form 180 monthly data. This subsystem uses Microsoft Excel spreadsheet to store, record, and update data. Historical data are presented in time series chart based on the data collection.

Both cases study were considered having equal number of incidences. However, the time series showed different trend. The trends were inspected through Excel where a constant trend is observed for Salmonellosis (Figure 4.1) and a downward trend for Tuberculosis (Figure 5.1). The trend results are important in order to determine the type of forecasting methods to be used in the model base management subsystem.

6.2 Analysis of Model Base Management Subsystem

Variation in the time series behavior was considered as one of causal factor on the difference in prediction results. This variation provides both insights into time series behavior and correlation with forecasting performance.

The overall results of model base management subsystem have been presented in Chapter 4 and Chapter 5. Table 6.1 looks at both cases to compare and contrast the application of the forecasting approach and the results.

Table 6.1 Summary of Forecasting Results

Zoonosis **Methods**	**Salmonellosis**	**Tuberculosis**
	Additive Model	**Multiplicative Model**
Moving Average	Formula: Eq. (3.1)	
	Result : Table 4.1 Figure 4.10	Result : Table 5.1 Figure 5.10
Regression	Formula: Eq. (3.2) and (3.3)	
	Result : Eq. (4.1) Table 4.3	Result : Eq. (5.1) Table 5.2
Decomposition	Additive decomposition Formula: Eq. (4.2) – (4.5), (4.7) Result : Eq. (4.6), (4.8), Table 4.3, Figure 4.3, Figure 4.4	Multiplicative decomposition Formula: Eq. (5.2) – (5.5), (5.7) Result : Eq. (5.6), (5.8), Table 5.3, Figure 5.3 Figure 5.4
Holt-Winter's	Additive Holt-Winter's Method: Eq. (3.5) – (3.8) Result : Table 4.5, α = 0.07, β = 0.01, γ = 0.01	Multiplicative Holt-Winter's Method: Eq. (3.9) – (3.12) Result : Table 5.4, α = 0.11, β = 0.01, γ = 0.01
ARIMA	Method: Eq. (3.13) – (3.18)	
	Result : Eq. (4.9) – (4.17) Figure 4.5 – 4.8 Table 4.6 ARIMA(9,0,14)(12,1,24)$_{12}$	Result : Eq. (5.9) – (5.18) Figure 5.5 – 5.8 Table 5.5 ARIMA(29,0,29)(12,1,12)$_{12}$
Neural Network	Method: Eq. (3.23)	
	Result : Table 4.8, Figure 4.9	Result : Table 5.7, Figure 5.9

Table 6.1 summarizes the forecasting methods applied in each case study, where Salmonellosis is showing a constant trend and Tuberculosis a downward trend. For each case study, the result of every method is reported. From Table 6.1, each zoonosis has different forecasting models. Even though some differences have been identified, some similarities could still be observed in the methods. It can be seen that for four methods (regression, moving average, ARIMA, and neural network), the same basic formula has been used, either for Salmonellosis or Tuberculosis. On the other hand,

for the other methods (decomposition and Holt-Winter's) different formulas have been applied. Additive decomposition and additive Holt-Winter's have been applied to Salmonellosis since it has an additive model because of its constant trend. In contrast, multiplicative decomposition and multiplicative Holt-Winter's have been applied to Tuberculosis~~,~~ because it was identified as multiplicative due to the downward trend of time series [104]

The results of different forecasting results for each case study were compared to identify the differences among them. When the differences were identified, Duncan multiple range test was conducted to determine which method will yield the difference. Duncan multiple range test was also able to show the suitable forecasting methods for each case study. In this study, the Duncan Multiple Range Test propose more than one suitable method. Thus, in this situation, the calculation for Coefficient of Variation (CV) was performed among the suitable methods to obtain the fittest methods.

Table 6.2 Summary of ANOVA, Duncan Test, and CV Results

Zoonosis	**Salmonellosis**	**Tuberculosis**
Methods	**Additive Model**	**Multiplicative Model**
ANOVA	• H_o was rejected • There were significant differences between forecasting methods	• H_o was rejected • There were significant differences between forecasting methods
Duncan Test	• *Appropriate*: Moving average Regression Holt-Winter's ARIMA Neural network • *Not appropriate:* Decomposition	• *Appropriate*: Holt-Winter's ARIMA Neural network • *Not appropriate*: Moving average Decomposition Regression
CV	The best forecasting method: regression analysis	The best forecasting method: neural network

Table 6.2 summarizes the specific results of ANOVA, Duncan multiple range test, and CV at individual case study. In both case studies, ANOVA has rejected the null hypothesis. This indicates that there are differences in the forecasting results from the actual data. Once ANOVA results have been obtained, the Duncan multiple range test was conducted to identify the method that suits the data. Even though, ANOVA yielded similar results for both case studies, the Duncan test produced different results. For Salmonellosis, only decomposition method is found not suitable while the other methods can be applied to the data. This was demonstrated by the Duncan results in Section 4.2.9. Decomposition mean, compared with actual data and other methods, was greater than the least significant range. For Tuberculosis, the detailed results were explained in Section 5.2.9. Duncan multiple range test revealed that moving average, decomposition, and regression are not suitable to be used for the existing data. Thus, Holt-Winter's, ARIMA, and neural network were appropriate for the Tuberculosis data.

The multiple forecasting methods were used to predict both case studies. However, there is still a need to obtain the fittest method. In order to fulfill this purpose, CV values for the suitable methods were calculated. The results yielded regression analysis as the fittest method for Salmonellosis case and neural network as the fittest method for Tuberculosis.

6.3 Analysis of Dialog Generation and Management Subsystem

In this research, DGMS components focused on the application of "What If Analysis", while the DSS interface is discussed solely in the separated subsection. The "What-If" (sensitivity) analysis was performed on the forecasting results before and after the addition of 2007 data to understand the effect of new data to the results in every forecasting method. The "What-If" analysis goals of both cases were similar. It is used to identify which forecasting method yielded the highest fluctuation along with the changing of data. The result of "what if" analysis for both case studies, which has been described earlier in section 4.4 and 5.4, is summarized in Table 6.3 and Table 6.4 for Salmonellosis and Tuberculosis, respectively. Based on the data, a sensitivity chart has been plotted for each case as shown in Figure 6.1 for Salmonellosis and Figure 6.2 for Tuberculosis.

Table 6.3 and Table 6.4 present the results for what-if analysis of all forecasting methods. The results of all methods are tabulated in one table for comparison and further observation. For each method, the difference of sensitivity analysis for 2008 (refer to Scenario 1 in Section 4.4 and 5.4) and the difference of sensitivity analysis for 2009 (refer to Scenario 2 in Section 4.4 and 5.4) are presented side by side. Fluctuations in each of the forecasting results can be clearly seen in the charts shown in Figure 6.1 for Salmonellosis and Figure 6.2 for Tuberculosis.

As shown in Table 6.3, the Salmonellosis results indicate a slight change of sensitivity difference both for 2008 and 2009: regression analysis (-1.84% to -0.63%), ARIMA (-0.96% to 3.04%), Holt-Winter's (1.35% to 6.91%), moving average (-8.58% to 7.71%), and decomposition (-9.22% to -0.22%). It is observed that moving average and decomposition methods exhibit greater sensitivity than the other three methods. The result obtained by neural network is unexpected where a large fluctuation in the data values is observed: 21.04 % to -38.59 % for 2008 and 15.89% to -40.19% for 2009. It is also shown in Figure 6.1 where the largest fluctuation in the forecasted data achieves by neural network both for 2008 and 2009. Also, a fluctuation spike, which occurs in December, can be observed in the results forecasted by neural network. The next largest fluctuation after neural network is shown by the decomposition results, both for 2008 and 2009 forecasts.

Different from Salmonellosis, the results in Table 6.4 for Tuberculosis indicate higher fluctuations in the forecasted values for 2008 and 2009 than the Salmonellosis results. The smallest fluctuation is achieved by Holt-Winter's (2.77% to 5.3%) followed by moving average (-0.31% to 6.35%) then ARIMA (-4.48% to 14.46%), and neural network (-33.57% to 17.34%). Decomposition method exhibits a large fluctuation from -179% to -71.9% while the largest fluctuations is observed in the results by regression analysis from -315% to 149.85%. Referring to Figure 6.2, the largest spike is shown by regression data in January 2008 and 2009, while the rest of the months in 2008 and 2009 indicate small fluctuations.

Table 6.3 Sensitivity Percentage of Salmonellosis

Mth	**Regression**		**Moving Average**		**Decomposition**		**Holt-Winter’s**		**ARIMA**		**Neural Network**	
	Forecast 2008	**Forecast 2009**	**Forecast 2008**	**Forecast 2009**	**Forecast 2008**	**Forecast 2009**	**Forecast 2008**	**Forecast 2009**	**Forecast 2008**	**Forecast 2009**	**Forecast 2008**	**Forecast 2009**
	Sens. (%)	**Sens. (%)**	**Sens. (%)**	**Sens. (%)**	**Sens. (%)**	**Sens. (%)**	**Sens. (%)**	**Sens. (%)**	**Sens. (%)**	**Sens. (%)**	**Sens. (%)**	**Sens. (%)**
Jan.	-1.84	-1.84	7.71	1.88	-7.4	-5.61	6.43	6.91	0.86	0.86	**21.04**	**15.89**
Feb.	-1.76	-1.75	3.58	-0.6	-9.22	-7.5	4.89	5.4	-0.96	-0.95	**-7.22**	**-12.93**
Mar.	-1.48	-1.47	3.36	-0.47	-6.38	-4.99	4.8	5.23	2.35	2.35	**-0.39**	**-3.76**
Apr.	-1.37	-1.37	0.21	-2.4	-7.69	-6.4	3.34	3.77	0.83	0.83	**14.26**	**15.04**
May	-1.13	-1.13	0.86	-2.46	-5.79	-4.74	3.03	3.41	-0.61	-0.61	**-3.35**	**6.26**
Jun.	-0.92	-0.92	2.1	-1.55	-2.83	-2.01	3.78	4.07	3.04	3.04	**18.36**	**13.19**
Jul.	-0.69	-0.69	-2.55	-4.07	-2.11	-1.49	1.68	1.93	-0.08	-0.08	**-0.97**	**6.1**
Aug.	-0.63	-0.63	-1.57	-4.57	-3.24	-2.68	1.5	1.73	0.09	0.09	**-10.33**	**-17.03**
Sept.	-0.67	-0.66	2.42	-3.23	-0.81	-0.22	3.84	4.06	-0.38	-0.37	**-13.54**	**-11.05**
Oct.	-0.71	-0.71	-3.14	-6.36	-4.18	-3.55	1.35	1.61	-0.19	-0.19	**18.56**	**12.65**
Nov.	-0.93	-0.93	-8.52	-8.58	-4.89	-4.04	2.14	2.46	0.09	0.09	**-9.2**	**-20.81**
Dec.	-0.69	-0.69	-7.67	-7.67	-1.43	-0.82	3.33	3.57	-0.26	-0.26	**-38.59**	**-40.19**

Table 6.4 Sensitivity Percentage of Tuberculosis

Mth	Regression		Moving Average		Decomposition		Holt-Winter’s		ARIMA		Neural Network	
	Forecast 2008	Forecast 2009	Forecast 2008	Forecast 2009	Forecast 2008	Forecast 2009	Forecast 2008	Forecast 2009	Forecast 2008	Forecast 2009	Forecast 2008	Forecast 2009
	Sens. (%)	Sens. (%)	Sens. (%)	Sens. (%)	Sens. (%)	Sens. (%)	Sens. (%)	Sens. (%)	Sens. (%)	Sens. (%)	Sens. (%)	Sens. (%)
Jan.	**149.83**	**-315.38**	3.28	4.77	-71.9	-108.35	4.72	5.3	5.61	6.59	-14.09	-9.27
Feb.	**11.82**	**15.13**	-0.31	3.32	-78.06	-117.08	3.37	3.97	6.49	6.93	-33.57	-33.27
Mar.	**5.6**	**6.75**	-0.3	3.45	-81.15	-122.21	3.1	3.7	2.19	2.52	0.66	1.07
Apr.	**4.59**	**5.55**	0	3.59	-84.41	-127.68	2.77	3.39	-3.98	-3.68	-10.03	-7.92
May	**5.52**	**6.52**	1.81	4.73	-85.41	-130.48	3.81	4.42	9.87	10.45	8.68	10.78
Jun.	**4.14**	**4.91**	1.59	4.76	-89.22	-136.9	3.37	4	-1.79	-1.57	-5.96	-11.45
Jul.	**5.1**	**6.06**	2.2	5.28	-91.81	-141.99	3.54	4.17	-3.14	-3.04	-15.44	-19.89
Aug.	**4.87**	**5.77**	2.79	5.71	-93.54	-146.14	4.23	4.86	5.05	5.39	-7.48	-8.09
Sept.	**6.31**	**7.45**	3.76	6.35	-98.6	-154.73	3.54	4.18	13.97	14.46	-1.13	-5.43
Oct.	**5.14**	**6.11**	3.17	5.76	-101.91	-161.33	3.62	4.27	6.02	6.39	15.48	17.34
Nov.	**7.19**	**8.53**	4.12	6.3	-104.48	-167.21	4.28	4.93	-4.48	-4.13	-3.59	-7.64
Dec.	**-0.02**	**0.16**	2.53	2.37	-111.7	-179.5	3.11	3.78	5.71	5.86	4.74	1.5

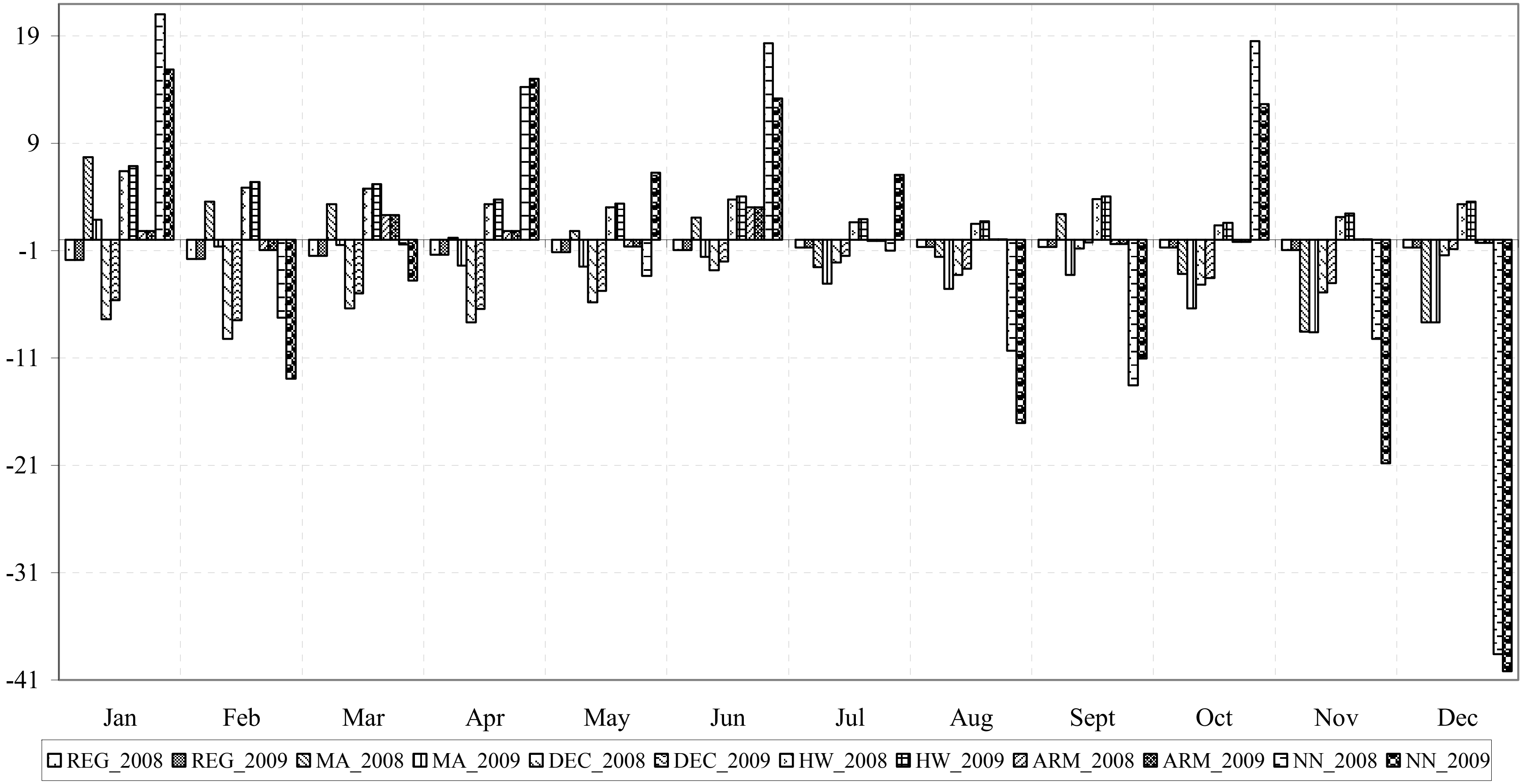


Figure 6.l Sensitivity Chart of Salmonellosis

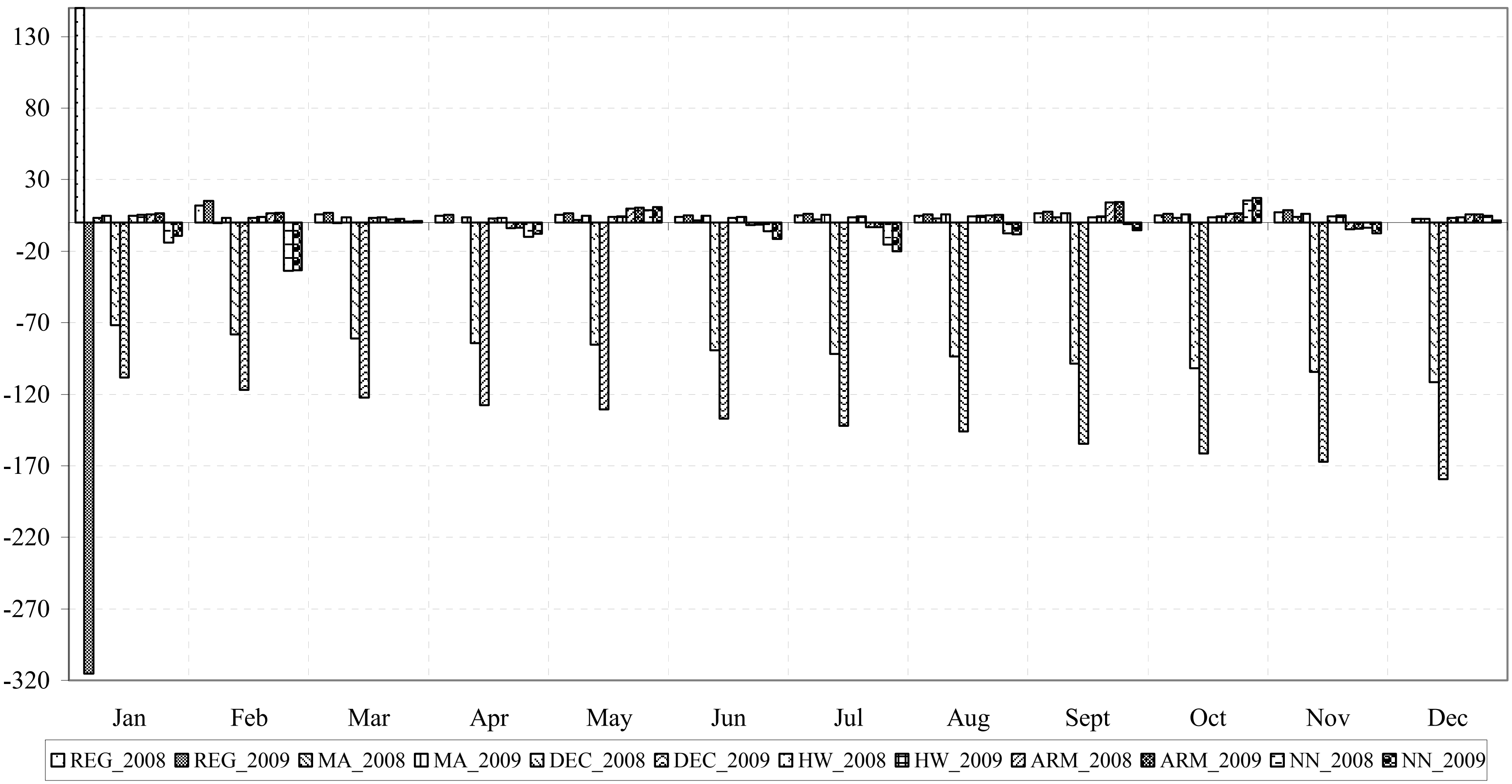


Figure 6.2 Sensitivity Chart of Tuberculosis

6.4 Cross Case Analysis

A comparative analysis has been conducted between the findings of Case 1 and Case 2. The results of each DSS components have been described in Section 6.2 to 6.4. The analysis has highlighted the similarities and differences between the two case studies. Based on the analysis of sensitivity difference in each case, a causal factor could be identified.

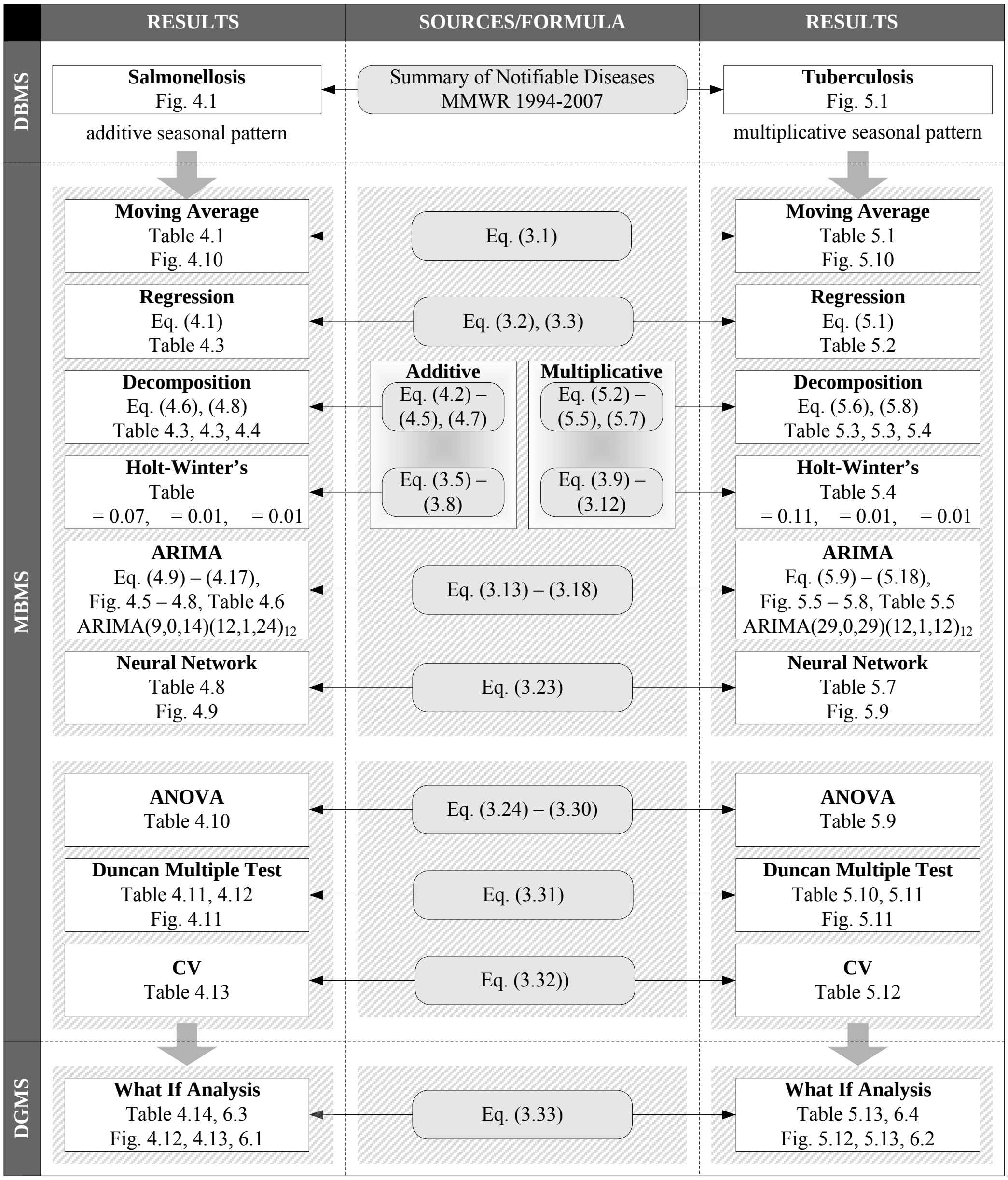


Figure 6.3 Summary of DSS Result

Figure 6.3 depicts a summary of the results associated with the method or formula used in the framework design. The different results were shown in the MBMS and DGMS. The CV results from the MBMS yield different results based on their forecasting method. This condition might happen because the application of specific forecasting results was dependent on the type of time series. The different fluctuations of "What-If" analysis between cases also proved that the prediction results associated with the number of time series and methods used. The main key to find the suitable forecasting method is by matching the model with historical data pattern as stated by Bowerman and O'Connell [104] that "*no single best forecasting model exists*". Since the Salmonellosis and Tuberculosis time series have different patterns, the results of the forecasting methods can be expected to be different too.

The DSS used in Case 1 and Case 2 have the same basic system structure, which were formed on monthly basis and occurred in the same country. US is a four-season country, where the type of diseases may differ with tropical country. This difference may cause the variation of time series trend. This research proposed the general framework that employs multi forecasting method that was able to process different types of time series. Using several available methods, there is a possibility in selecting the most appropriate method that produces the better results. Besides the forecasting method, the changes in the time series may yield the different results. It was shown by sensitivity analysis results. Using this analysis, user could identify the result fluctuations in different method. The results provided the user which forecasting method was more stable based on data updated. Couple with ANOVA, Duncan Test and CV results, user could choose appropriate method that having the smallest variation and relatively produced little fluctuation triggered by the changing of data.

The features of the proposed framework were more comprehensive than the existing model as reviewed in the related works. The comparison between proposed framework and previous model is listed in Table 6.5. As can be seen from Table 6.5, the proposed model has DSS components, including DBMS, MBMS, and DGMS. These features are more complete than other related works, which are supported by database management subsystem and model base management subsystem only.

Table 6.5 Feature Comparison between Related Works and Proposed Framework

DSS Components Researches	DBMS		MBMS		DGMS
	Single Disease	**Multi Disease**	**Single Method**	**Multi Method**	
Luz *et al.*, 2008; Thammapalo *et al.*, 2005; Chaves and Pascual, 2006; Chaves and Pascual, 2007; Gomez-Elipe *et al.*, 2007; Rakotomanana et al., 2007; Teklehaimanot *et al.*, 2004; Konchom et al., 2006; Daniel *et al.*, 2007; Maijala and Ranta, 2003; Lai, 2005; Shtatland *et al.*, 2006; Earnest *et al.*, 2005; Sebastiani *et al.*, 2006; Beck-Wörner *et al.*, 2007; Medina *et al.*, 2008; Hammad *et al.*, 1996; Debanne et al., 2000; Naumova *et al.*, 2005; Tanner *et al.*, 2008	√		√		NA
Shankar et al., 2003; Sai *et al.*, 2004; Briët et al., 2008	√			√	NA
Proposed Framework		√		√	√

6.5 Graphical User Interface (GUI) Design

This subchapter describes the logic used to create the VBA for Excel model and also includes the different worksheets. This subchapter also explains the interface that integrates 3 DSS components in a single DSS interface.

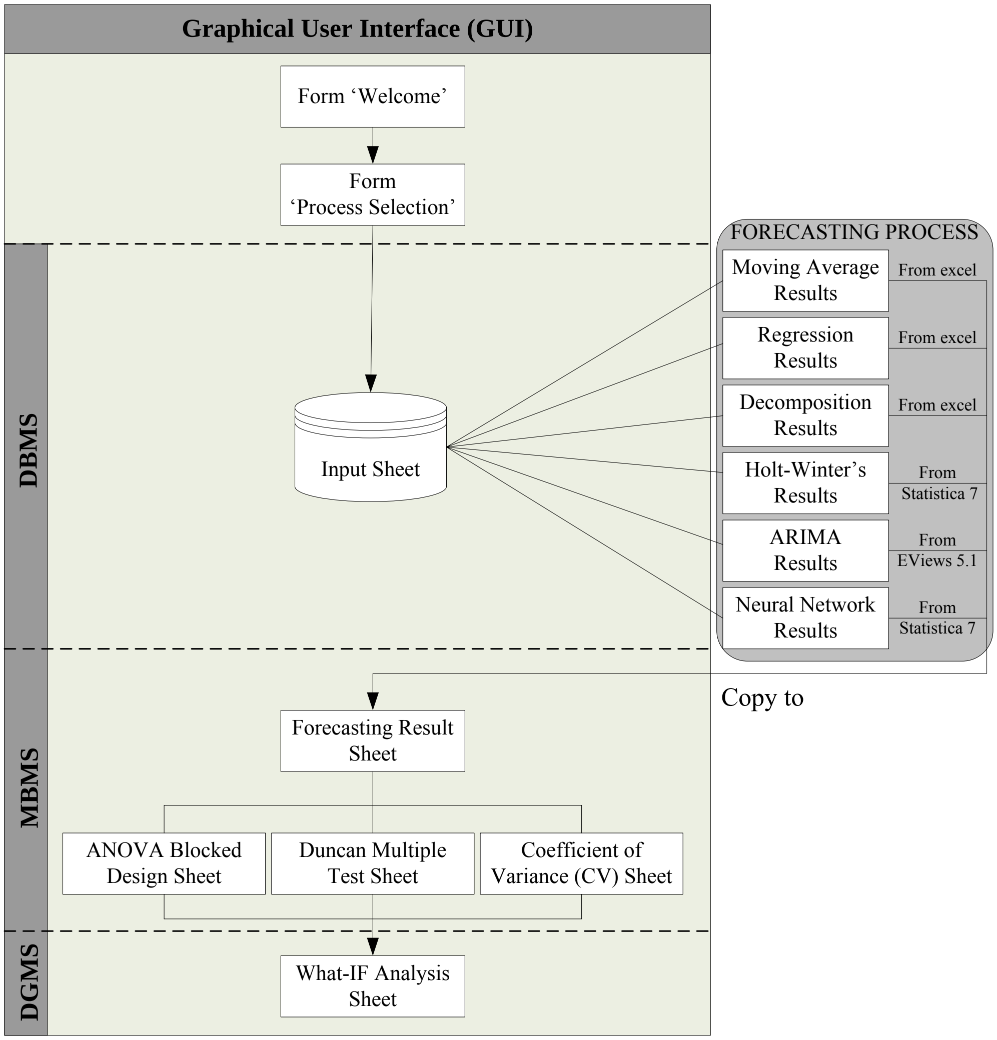


Figure 6.4 GUI Flow of the Proposed Framework

Figure 6.4 illustrates the flow of DSS GUI based on the proposed framework. The GUI is opened by "Welcome" form. It is followed by "Process Selection" form. Through this form, three DSS components are presented and can be selected by the user. As seen in Figure 6.4, different software is used for processing forecasting results for each method within the MBMS component. Then, the results are copied into the excel sheet to be further processed.The detail interface design for each process is presented on the next section of this subchapter.

The system was designed using spreadsheet based DSS. The application was opened by “Welcome” form as seen in Figure 6.5. This form consists of two command button, namely “About” and “ENTER”. Button “About” display application description (Figure 6.6), while when user click button “ENTER” then it open the application by selection of process in Figure 6.7.

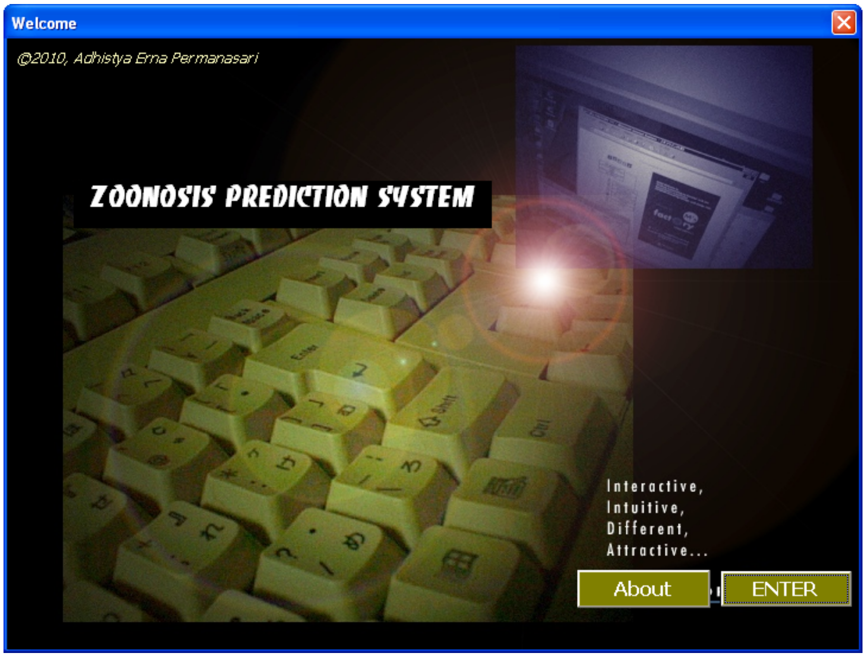
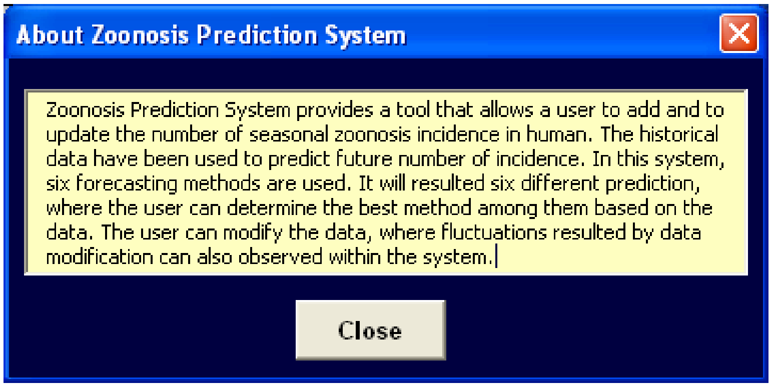


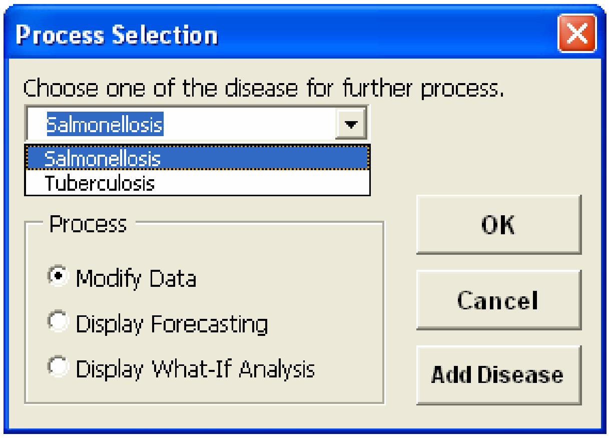
Figure 6.5 Welcome Form



Zoonosis Prediction System provides a tool that allows a user to add and to update the number of seasonal zoonosis incidence in human. The historical data have been used to predict future number of incidence. In this system, six forecasting methods are used. It will resulted six different prediction, where the user can determine the best method among them based on the data. The user can modify the data, where fluctuations resulted by data modification can also observed within the system.

Figure 6.6 About Form

User form “Process Selection” (Figure 6.7) is divided into two parts, drop down box for disease selection and frame “Process” for process selection. Using “Process” users can choose which component of DSS will be processed. There are three radio buttons to represents each DSS component: “Modify Data” for DBMS, “Display Forecasting” for MBMS and “Display What-If Analysis for DGMS.



Process Selection

Choose one of the disease for further process.

Salmonellosis

Salmonellosis
Tuberculosis

Process

● Modify Data
○ Display Forecasting
○ Display What-If Analysis

Figure 6.7 Process Selection Form

Within this form, user can also add more disease using command button “Add Disease”. In Figure 6.7, Salmonellosis is chosen as the case study for the following interfaces.

Spreadsheet DSS was applied for the zoonosis prediction application. Therefore, there are several sheets in DSS application. Figure 6.8 shows input sheets for Salmonellosis dataset. Input table is divided into three columns: t, Month, and No of Incidence.

Input spreadsheet is completed by 3 command button: Continue (for next process), Return to Process Selection (for back to Process Selection form in Figure 6.7), and End (for closing the sheet).

Once data are entered to the sheet, forecasting results can be calculated. The data are processed through different tools and results for each method are copied into Forecasting Results sheet as seen in Figure 6.9.

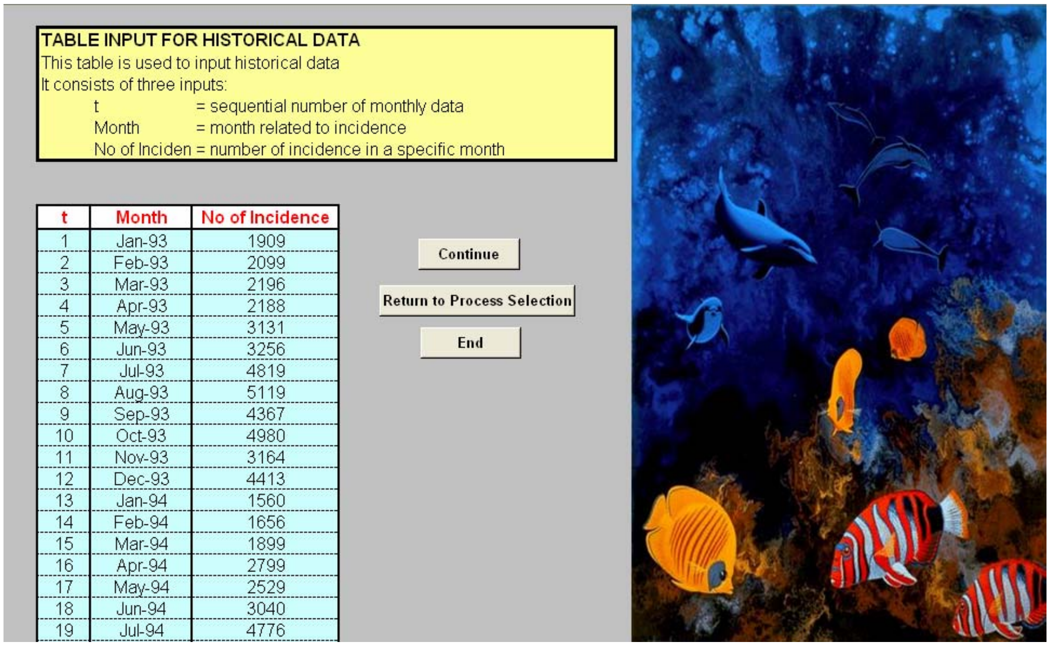
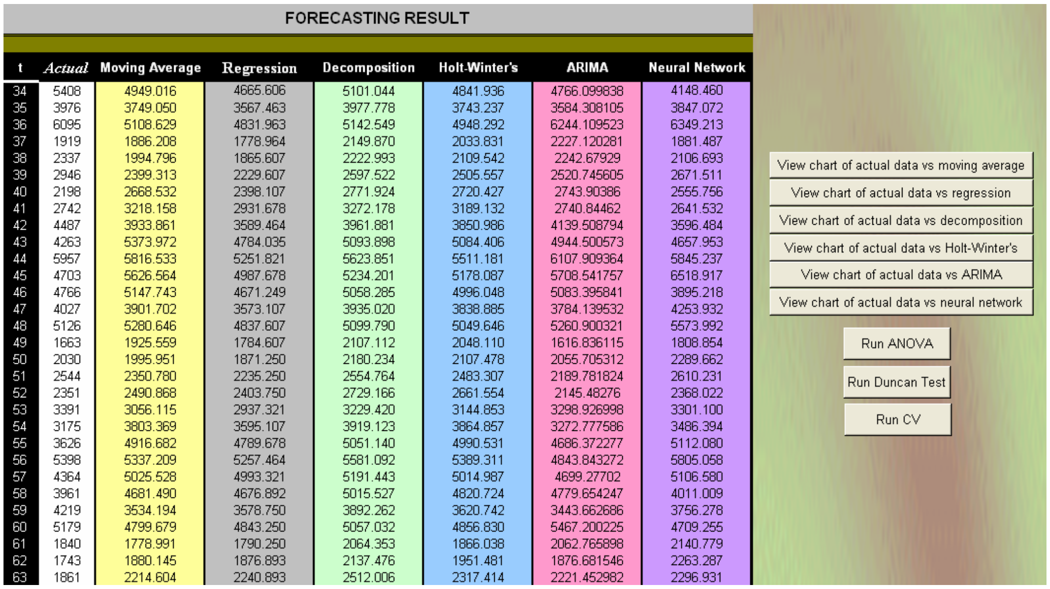


TABLE INPUT FOR HISTORICAL DATA
This table is used to input historical data
It consists of three inputs:
t = sequential number of monthly data
Month = month related to incidence
No of Inciden = number of incidence in a specific month

t	Month	No of Incidence
1	Jan-93	1909
2	Feb-93	2099
3	Mar-93	2196
4	Apr-93	2188
5	May-93	3131
6	Jun-93	3256
7	Jul-93	4819
8	Aug-93	5119
9	Sep-93	4367
10	Oct-93	4980
11	Nov-93	3164
12	Dec-93	4413
13	Jan-94	1560
14	Feb-94	1656
15	Mar-94	1899
16	Apr-94	2799
17	May-94	2529
18	Jun-94	3040
19	Jul-94	4776

Figure 6.8 Input Table



FORECASTING RESULT

t	*Actual*	Moving Average	Regression	Decomposition	Holt-Winter's	ARIMA	Neural Network
34	5408	4949.016	4665.606	5101.044	4841.936	4766.099838	4148.460
35	3976	3749.050	3567.463	3977.778	3743.237	3584.308105	3847.072
36	6095	5108.629	4831.963	5142.549	4948.292	6244.109523	6349.213
37	1919	1886.208	1778.964	2149.870	2033.831	2227.120281	1881.487
38	2337	1994.796	1865.607	2222.993	2109.542	2242.67929	2106.693
39	2946	2399.313	2229.607	2597.522	2505.557	2520.745605	2671.511
40	2198	2668.532	2398.107	2771.924	2720.427	2743.90386	2555.756
41	2742	3218.158	2931.678	3272.178	3189.132	2740.84462	2641.532
42	4487	3933.861	3589.464	3961.881	3850.986	4139.508794	3596.484
43	4263	5373.972	4784.035	5093.898	5084.406	4944.500573	4657.953
44	5957	5816.533	5251.821	5623.851	5511.181	6107.909364	5845.237
45	4703	5626.564	4987.678	5234.201	5178.087	5708.541757	6518.917
46	4766	5147.743	4671.249	5058.285	4996.048	5083.395841	3895.218
47	4027	3901.702	3573.107	3935.020	3838.885	3784.139532	4253.932
48	5126	5280.646	4837.607	5099.790	5049.646	5260.900321	5573.992
49	1663	1925.559	1784.607	2107.112	2048.110	1616.836115	1808.854
50	2030	1995.951	1871.250	2180.234	2107.478	2055.705312	2289.662
51	2544	2350.780	2235.250	2554.764	2483.307	2189.781824	2610.231
52	2351	2490.868	2403.750	2729.166	2661.554	2145.48276	2368.022
53	3391	3056.115	2937.321	3229.420	3144.853	3298.926998	3301.100
54	3175	3803.369	3595.107	3919.123	3864.857	3272.777586	3486.394
55	3626	4916.682	4789.678	5051.140	4990.531	4686.372277	5112.080
56	5398	5337.209	5257.464	5581.092	5389.311	4843.843272	5805.058
57	4364	5025.528	4993.321	5191.443	5014.987	4699.27702	5106.580
58	3961	4681.490	4676.892	5015.527	4820.724	4779.654247	4011.009
59	4219	3534.194	3578.750	3892.262	3620.742	3443.662686	3756.278
60	5179	4799.679	4843.250	5057.032	4856.830	5467.200225	4709.255
61	1840	1778.991	1790.250	2064.353	1866.038	2062.765898	2140.779
62	1743	1880.145	1876.893	2137.476	1951.481	1876.681546	2263.287
63	1861	2214.604	2240.893	2512.006	2317.414	2221.452982	2296.931

View chart of actual data vs moving average
View chart of actual data vs regression
View chart of actual data vs decomposition
View chart of actual data vs Holt-Winter's
View chart of actual data vs ARIMA
View chart of actual data vs neural network
Run ANOVA
Run Duncan Test
Run CV

Figure 6.9 Forecasting Results Sheet

The actual data and forecasting results are presented in this sheet, where the user can select from the list of command button to display a specific chart. If the user selects the first option, then the chart is shown graphically as in Figure 6.10 and the selection of the last option illustrates the chart in Figure 6.11. In each chart, there is a command button “View Forecasting Sheet” that allow the user for back to forecasting results sheet.

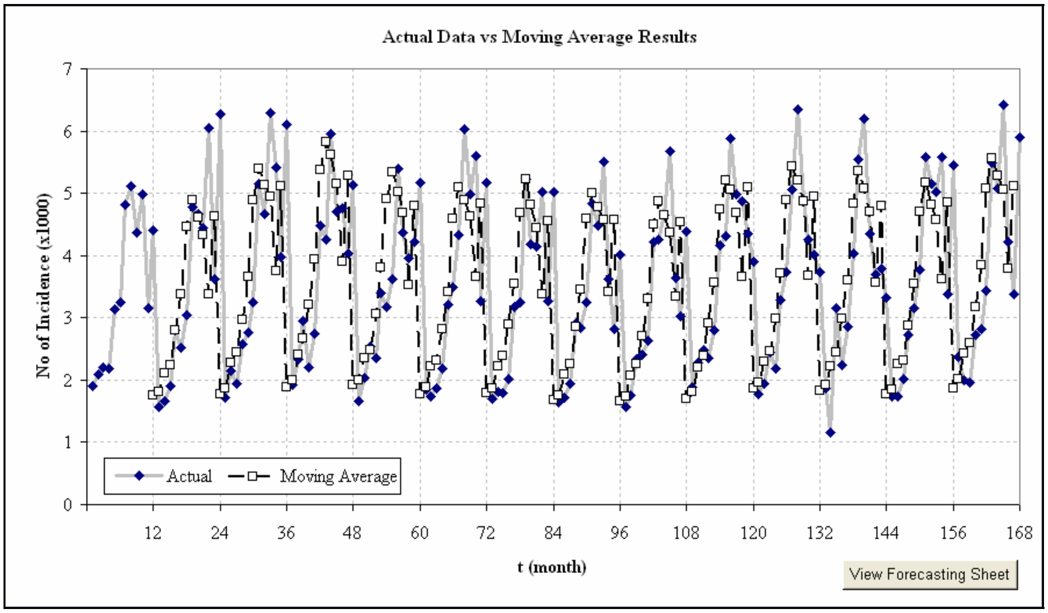


Figure 6.10 Actual Data vs Moving Average Chart

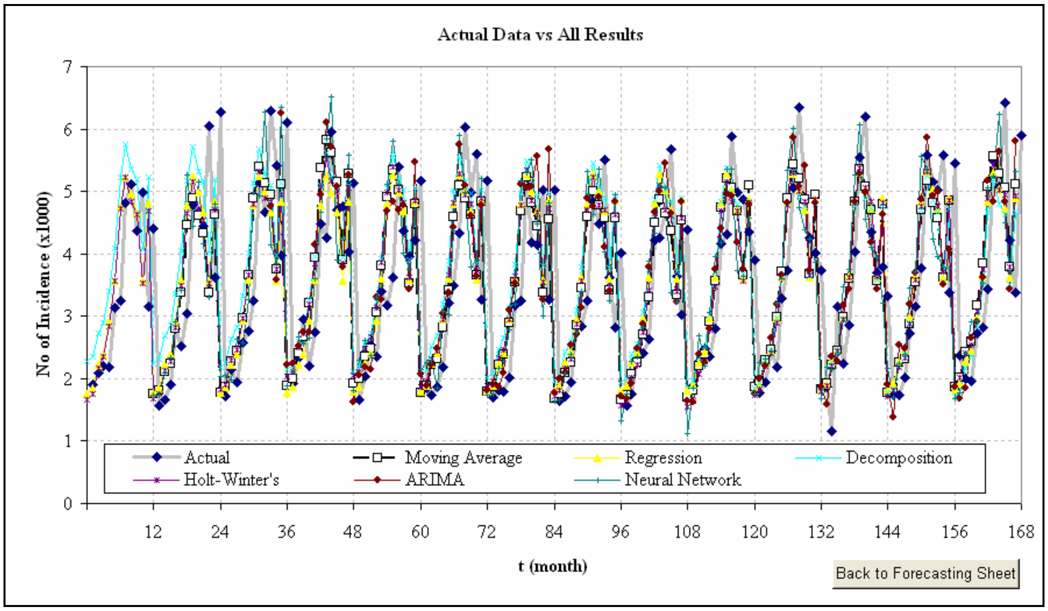
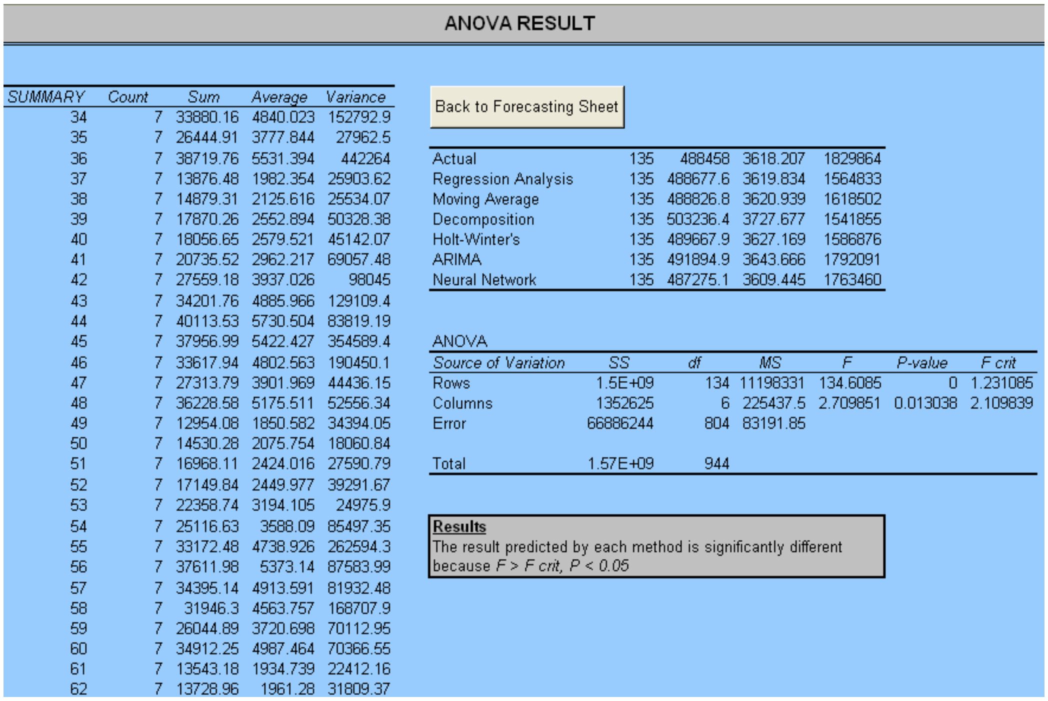


Figure 6.11 Actual Data vs All Results Chart

Figure 6.9 represents the process in the MBMS component where beside the radio button list, it is also provided others calculation in the MBMS, including ANOVA, Duncan Multiple Range Test, and CV. The process can be accessed through the respective command button. When the user click the buttons, “ANOVA Result” (Figure 6.12), “Duncan Multiple Range Test Result” (Figure 6.13), and “Coefficient of Variance” (Figure 6.14) are displayed sequentially.



ANOVA RESULT

SUMMARY	Count	Sum	Average	Variance
34	7	33880.16	4840.023	152792.9
35	7	26444.91	3777.844	27962.5
36	7	38719.76	5531.394	442264
37	7	13876.48	1982.354	25903.62
38	7	14879.31	2125.616	25534.07
39	7	17870.26	2552.894	50328.38
40	7	18056.65	2579.521	45142.07
41	7	20735.52	2962.217	69057.48
42	7	27559.18	3937.026	98045
43	7	34201.76	4885.966	129109.4
44	7	40113.53	5730.504	83819.19
45	7	37956.99	5422.427	354589.4
46	7	33617.94	4802.563	190450.1
47	7	27313.79	3901.969	44436.15
48	7	36228.58	5175.511	52556.34
49	7	12954.08	1850.582	34394.05
50	7	14530.28	2075.754	18060.84
51	7	16968.11	2424.016	27590.79
52	7	17149.84	2449.977	39291.67
53	7	22358.74	3194.105	24975.9
54	7	25116.63	3588.09	85497.35
55	7	33172.48	4738.926	262594.3
56	7	37611.98	5373.14	87583.99
57	7	34395.14	4913.591	81932.48
58	7	31946.3	4563.757	168707.9
59	7	26044.89	3720.698	70112.95
60	7	34912.25	4987.464	70366.55
61	7	13543.18	1934.739	22412.16
62	7	13728.96	1961.28	31809.37

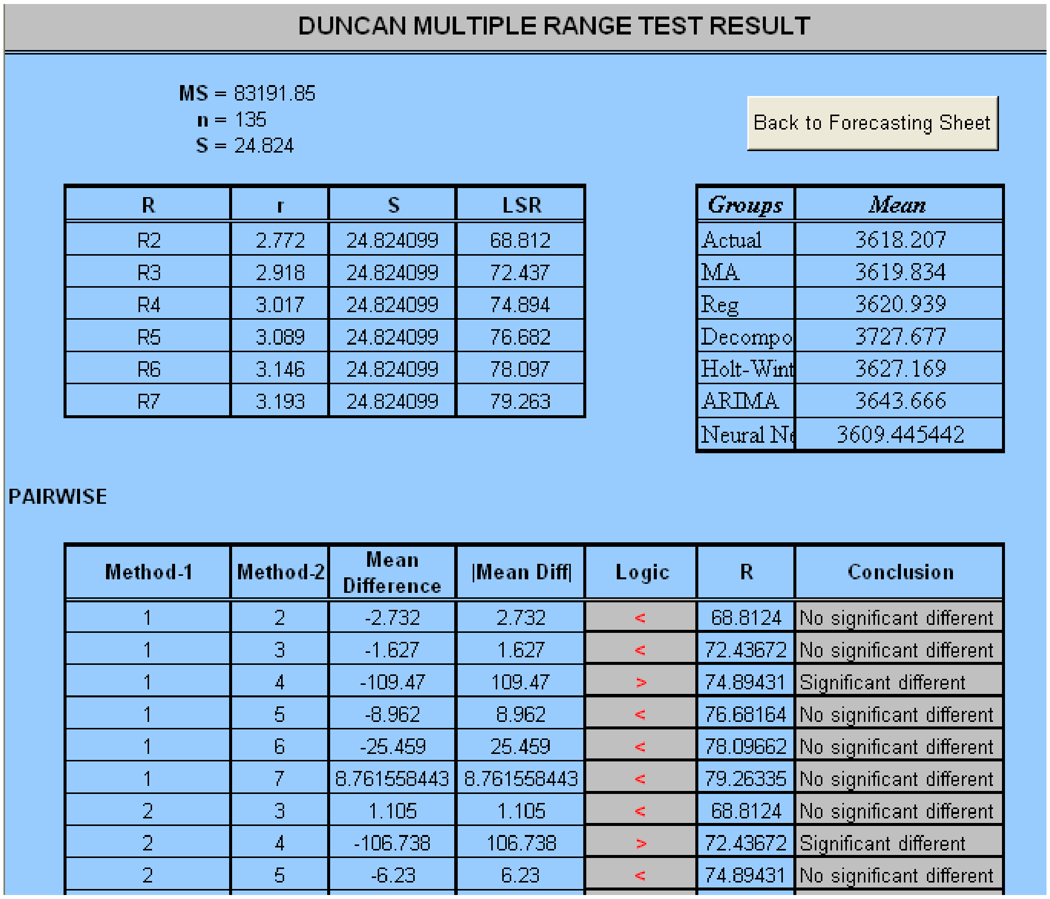
Actual	135	488458	3618.207	1829864
Regression Analysis	135	488677.6	3619.834	1564833
Moving Average	135	488826.8	3620.939	1618502
Decomposition	135	503236.4	3727.677	1541855
Holt-Winter's	135	489667.9	3627.169	1586876
ARIMA	135	491894.9	3643.666	1792091
Neural Network	135	487275.1	3609.445	1763460

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1.5E+09	134	11198331	134.6085	0	1.231085
Columns	1352625	6	225437.5	2.709851	0.013038	2.109839
Error	66886244	804	83191.85			
Total	1.57E+09	944				

Results
The result predicted by each method is significantly different because *F > F crit, P < 0.05*

Figure 6.12 ANOVA Result



DUNCAN MULTIPLE RANGE TEST RESULT

MS = 83191.85
n = 135
S = 24.824

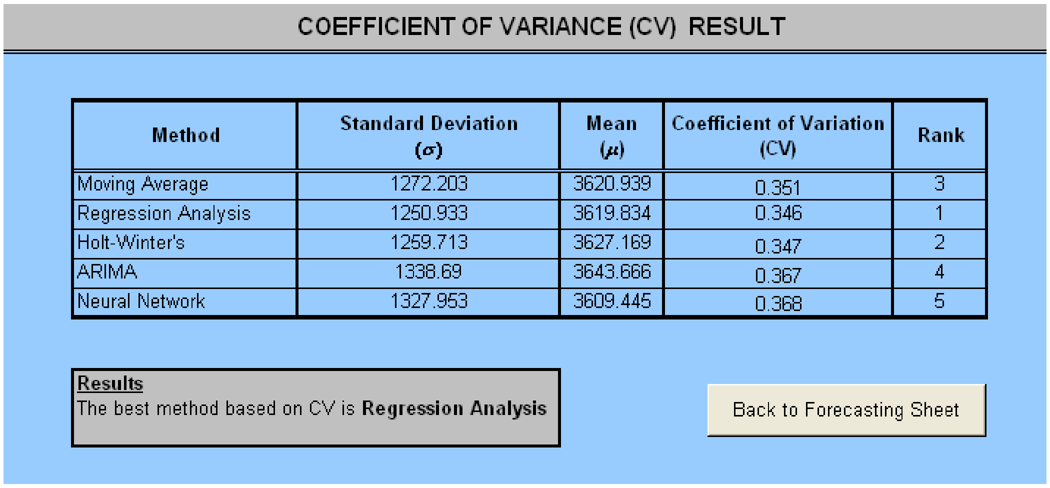
R	r	S	LSR
R2	2.772	24.824099	68.812
R3	2.918	24.824099	72.437
R4	3.017	24.824099	74.894
R5	3.089	24.824099	76.682
R6	3.146	24.824099	78.097
R7	3.193	24.824099	79.263

Groups	*Mean*
Actual	3618.207
MA	3619.834
Reg	3620.939
Decompo	3727.677
Holt-Wint	3627.169
ARIMA	3643.666
Neural Ne	3609.445442

PAIRWISE

Method-1	Method-2	Mean Difference	|Mean Diff|	Logic	R	Conclusion
1	2	-2.732	2.732	<	68.8124	No significant different
1	3	-1.627	1.627	<	72.43672	No significant different
1	4	-109.47	109.47	>	74.89431	Significant different
1	5	-8.962	8.962	<	76.68164	No significant different
1	6	-25.459	25.459	<	78.09662	No significant different
1	7	8.761558443	8.761558443	<	79.26335	No significant different
2	3	1.105	1.105	<	68.8124	No significant different
2	4	-106.738	106.738	>	72.43672	Significant different
2	5	-6.23	6.23	<	74.89431	No significant different

Figure 6.13 Duncan Results



COEFFICIENT OF VARIANCE (CV) RESULT

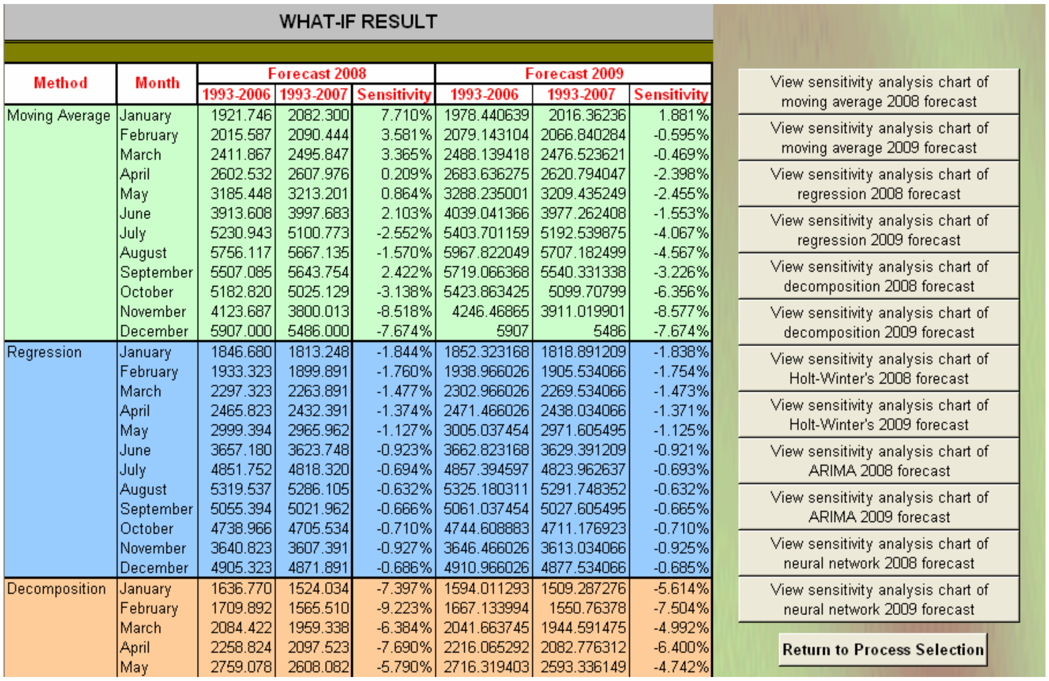
Method	Standard Deviation (σ)	Mean (μ)	Coefficient of Variation (CV)	Rank
Moving Average	1272.203	3620.939	0.351	3
Regression Analysis	1250.933	3619.834	0.346	1
Holt-Winter's	1259.713	3627.169	0.347	2
ARIMA	1338.69	3643.666	0.367	4
Neural Network	1327.953	3609.445	0.368	5

Results
The best method based on CV is **Regression Analysis**

Figure 6.14 Coefficient of Variance Result

When users have finished the process in the MBMS, they can back to the "Process Selection" form to continue accessing other processes. The last option is "Display

What-If Analysis”. The interface of this process is showed in Figure 6.15. The interface contains several buttons for navigating to the various chart sheets. The example of the first chart can be seen in Figure 6.16.



WHAT-IF RESULT

Method	Month	Forecast 2008			Forecast 2009		
		1993-2006	1993-2007	Sensitivity	1993-2006	1993-2007	Sensitivity
Moving Average	January	1921.746	2082.300	7.710%	1978.440639	2016.36236	1.881%
	February	2015.587	2090.444	3.581%	2079.143104	2066.840284	-0.595%
	March	2411.867	2495.847	3.365%	2488.139418	2476.523621	-0.469%
	April	2602.532	2607.976	0.209%	2683.636275	2620.794047	-2.398%
	May	3185.448	3213.201	0.864%	3288.235001	3209.435249	-2.455%
	June	3913.608	3997.683	2.103%	4039.041366	3977.262408	-1.553%
	July	5230.943	5100.773	-2.552%	5403.701159	5192.539875	-4.067%
	August	5756.117	5667.135	-1.570%	5967.822049	5707.182499	-4.567%
	September	5507.085	5643.754	2.422%	5719.066368	5540.331338	-3.226%
	October	5182.820	5025.129	-3.138%	5423.863425	5099.70799	-6.356%
	November	4123.687	3800.013	-8.518%	4246.46865	3911.019901	-8.577%
	December	5907.000	5486.000	-7.674%	5907	5486	-7.674%
Regression	January	1846.680	1813.248	-1.844%	1852.323168	1818.891209	-1.838%
	February	1933.323	1899.891	-1.760%	1938.966026	1905.534066	-1.754%
	March	2297.323	2263.891	-1.477%	2302.966026	2269.534066	-1.473%
	April	2465.823	2432.391	-1.374%	2471.466026	2438.034066	-1.371%
	May	2999.394	2965.962	-1.127%	3005.037454	2971.605495	-1.125%
	June	3657.180	3623.748	-0.923%	3662.823168	3629.391209	-0.921%
	July	4851.752	4818.320	-0.694%	4857.394597	4823.962637	-0.693%
	August	5319.537	5286.105	-0.632%	5325.180311	5291.748352	-0.632%
	September	5055.394	5021.962	-0.666%	5061.037454	5027.605495	-0.665%
	October	4738.966	4705.534	-0.710%	4744.608883	4711.176923	-0.710%
	November	3640.823	3607.391	-0.927%	3646.466026	3613.034066	-0.925%
	December	4905.323	4871.891	-0.686%	4910.966026	4877.534066	-0.685%
Decomposition	January	1636.770	1524.034	-7.397%	1594.011293	1509.287276	-5.614%
	February	1709.892	1565.510	-9.223%	1667.133994	1550.76378	-7.504%
	March	2084.422	1959.338	-6.384%	2041.663745	1944.591475	-4.992%
	April	2258.824	2097.523	-7.690%	2216.065292	2082.776312	-6.400%
	May	2759.078	2608.082	-5.790%	2716.319403	2593.336149	-4.742%

View sensitivity analysis chart of moving average 2008 forecast
View sensitivity analysis chart of moving average 2009 forecast
View sensitivity analysis chart of regression 2008 forecast
View sensitivity analysis chart of regression 2009 forecast
View sensitivity analysis chart of decomposition 2008 forecast
View sensitivity analysis chart of decomposition 2009 forecast
View sensitivity analysis chart of Holt-Winter's 2008 forecast
View sensitivity analysis chart of Holt-Winter's 2009 forecast
View sensitivity analysis chart of ARIMA 2008 forecast
View sensitivity analysis chart of ARIMA 2009 forecast
View sensitivity analysis chart of neural network 2008 forecast
View sensitivity analysis chart of neural network 2009 forecast
Return to Process Selection

Figure 6.15 What-If Result

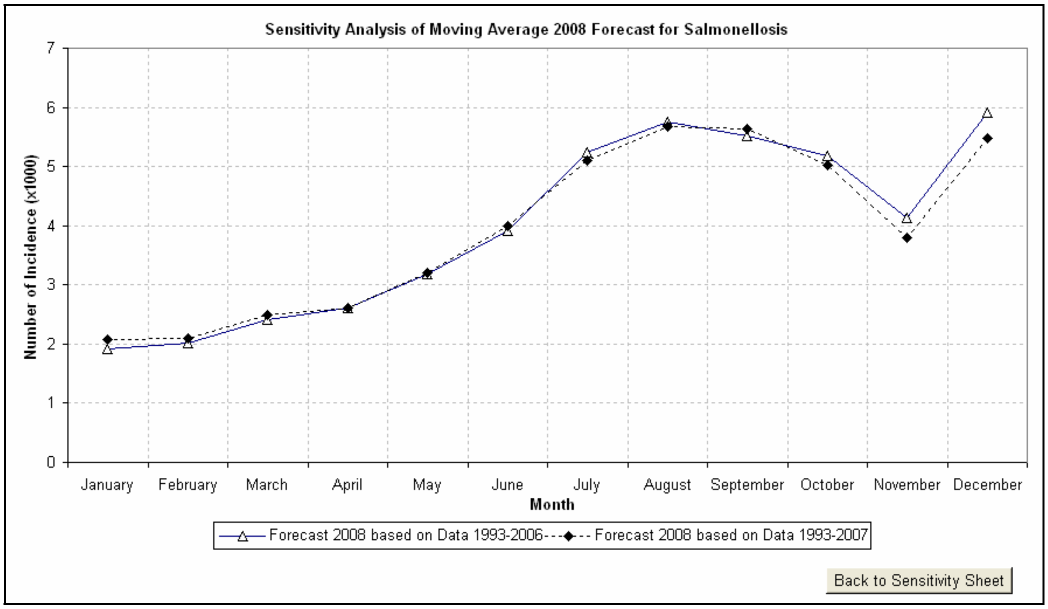


Figure 6.16 Chart of Sensitivity Analysis of Moving Average 2008 Forecast

6.6 Framework Benefit

The zoonosis DSS presents two predictions of human zoonosis based on seasonal pattern. Case study 1 reports DSS result on additive seasonal model using Salmonella as the selected zoonosis, while case study 2 reports DSS result on multiplicative seasonal model using Tuberculosis. The general zoonosis prediction framework has provided a systematic process forecasting future number of seasonal zoonosis, not only to the provided case studies but its use could also be extended to other seasonal zoonosis.

A number of general findings that have become evident with initial uses of the study:

- The proposed framework could be used to identify and recognize the main components to develop a DSS model for zoonosis prediction. It has been developed with a good understanding of the input, output, and how to design a system using a mathematical model.
- The proposed framework can be used as a decision support model. It has provided a basis for predicting seasonal zoonosis incidence from historical data.
- The application of the proposed framework came out with GUI design that was a representation of three DSS components into a user friendly interface. User can access the system through this interface.

6.7 Summary

The proposed DSS framework represented two case studies, namely Salmonellosis and Tuberculosis. These two cases had the same DSS components. However, each of them represented different time series model, where Salmonellosis was for additive time series model and Tuberculosis was for multiplicative time series model.

The results of applying both seasonality patterns were compared and discussed in this chapter. The analysis results and findings from both case studies confirmed the three DSS components. The application and results of the proposed DSS framework were varied for different cases. The cross case analysis provided the identifications of causal factor that influenced the differences between cases study results. The factor that affects the suitability of the model to be used in the forecast has been identified as due to the type of seasonality pattern of time series, either additive seasonality or multiplicative seasonality. The results indicate popular methods, such as ARIMA, may not yield the best result.

CHAPTER 7
CONCLUSION

7.1 Dissertation Summary

Widespread outbreak of zoonosis around the world coupled with the increasing number of incidence motivated a research on emerging zoonosis. This research provides a comprehensive study focused on zoonosis impact faced by community around the world. The main goal of the research was to develop a zoonosis system framework for assisting decision making and prediction of the future number of zoonosis incidence in human. Hence, it focused on seasonal zoonosis.

In this research, a DSS framework of seasonal zoonosis was developed to solve the problem stated in Chapter 1. The four methodology stages were considered in detail and used on the selected case studies (refer to Figure 3.6), namely background, model design, model development, and evaluation. The framework was divided into two seasonal model, additive seasonal model and multiplicative seasonal model. Criteria for selection of case studies were based on broad occurrences, high number of incidence, completeness and availability of data. Thus, Salmonellosis was selected to represent an additive seasonal model, while Tuberculosis was used to represent a multiplicative seasonal model.

The proposed DSS framework enabled the prediction of human incidence in seasonal zoonosis using different forecasting methods. Within MBMS, the framework provides a feature to determine the appropriate model using ANOVA and Duncan Multiple Range Test. In the DGMS, data could be changed by the user including adding new data, updating data, and deleting data. The resulting fluctuations in forecasting results were quantified using What-If analysis (sensitivity analysis). Refer to the What-If analysis results, the user could choose the method with the smallest fluctuation.

The proposed DSS framework was able to be applied into two time series model. This research shows that DSS can be developed to support zoonosis prediction. The framework presented in this research is hoped to give a contribution in the application for prediction of future number of human incidence in other seasonal zoonosis. As the conclusion, the framework assisted the user to choose the suitable forecasting method with a small fluctuation that triggered by data updated. Finally, the GUI design was provided as the interface connection between user and system.

7.2 Research Findings

The overall aim of this research is to develop a decision support system (DSS) framework that incorporates the use of multi forecasting techniques to predict seasonal zoonosis incidence of multi diseases. The framework could be used to predict two case studies, either of additive or multiplicative seasonal time series trends, using the appropriate models.

The research was conducted to answer the research questions and fulfills the objectives. The objectives were used as guidance for model development and analysis of the proposed framework. A number of findings were identified based on research objectives presented in Chapter 1. The findings based on the objectives are summarized in the sequel.

1. **Objective 1**: *To acquire information of previous and current research works dealing with zoonosis prediction system and to identify the research gap that needs to be addressed by the proposed model.*

   Findings: The research investigated and reviewed various related work of zoonosis. Reviews of literatures are presented not only related on zoonosis impact on human, but also on zoonosis researches in prediction system, especially to estimate future number of zoonosis in human. The literature on zoonosis prediction used at most only three forecasting methods and focused only on one type of disease. The related studies in Chapter 2 presented the prediction model with most of them using only single forecasting method (4 works from total of 23 works). It highlighted the importance of zoonosis prediction framework using multi forecasting methods for various diseases to obtain the most appropriate model with better result, whereas the framework was hoped to be applied in various diseases.

2. **Objective 2**: *To develop a DSS framework that can be applied to different seasonal zoonosis which covers two kinds of time series models, additive time series and multiplicative time series*.

   Findings: The research revealed a DSS framework on zoonosis prediction based on the research scope. The framework was a general framework that can be applied to different seasonal zoonosis as presented in Chapter 3. It covers two types of time series models within model base management system, additive time series and multiplicative time series. Additive time series is time series that exhibits constant fluctuation, while multiplicative time series is time series that presents the relative upward or downward trend. The framework presented a detailed description and graphical presentation of the development flow on model base management subsystem. Due to the difference in formula based on the data trend for some method, the framework consisted of two different forecasting approaches associated for each time series. Hence using ANOVA and Duncan Mutliple Range Test, the appropriate method among available approaches could be identified. Coefficient of Variation (CV) was conducted into the set of suitable methods to determine the best method. Finally, the framework provided a sensitivity analysis in the DGMS component in order to identify the fluctuation of results on all forecasting method that were influenced by the changing of the data.

3. **Objective 3**: *To apply each DSS component into the selected case studies*.

   Findings: The research developed the proposed framework into two different diseases, in which one was Salmonellosis to represent the additive time series and the other was Tuberculosis to represent multiplicative time series. The case studies were selected based on the availability of monthly data. To ease the discussion of result comparison between them, the zoonosis time series was collected from the same sources and has the same number of sample.

4. **Objective 4**: *To analyze the results comparison of both case studies and to develop GUI design based on DSS component.*

   Findings: The research reported the final results of both case studies and evaluates the results of both. The results were analyzed and reported to distinguish the differences and the similarities between them. The results indicate that popular methods, such as ARIMA, may not yield the best result. Once the comparison analysis was done, the GUI interface was design to connect the interaction between user and system.

7.3 Future Works

This research addressed the application of DSS model to predict number of human incidence caused by zoonosis. The results met its specified goal. The results also provided some contributions to the field of study. However, several aspects have been identified to expand the result of dissertation and also to improve the proposed framework. The following were some opportunities that could be conducted for research extension.

- Expansion of the proposed framework by designing DSS software based on the framework presented in this research.

- Six forecasting methods, including five statistical methods and one soft computing method, were used to develop the MBMS components. These methods were selected based on their varying model complexity and the numerical historical data were readily available, whereas all predictions were based on such data. Furthermore, more forecasting methods can be added within this component to improve the prediction results. For example: if the incomplete data are found, then different approach, such as Bayesian method, could be considered to solve this problem.

- Further study can be conducted for investigating and modeling a DSS framework for nonseasonal zoonosis by using the proposed methodology. Even though most of the stages may be similar, but the forecasting approach might be different. Different methods will have their own formula in handling seasonal and nonseasonal time series. Furthermore, the results can be analyzed and compared with the seasonal framework to identify the similarities and the differences between them.

- The scope of this research focuses on predicting human number of incidence. The proposed framework also can be extended as conceptual framework to be applied into animal incidence of zoonosis, whereas some zoonosis also attack animal, for example Swine Flu, Avian Influenza, Rabies, etc.

To summarize, Figure 7.1 illustrates the flow of research work that has been completed and direction of future works.

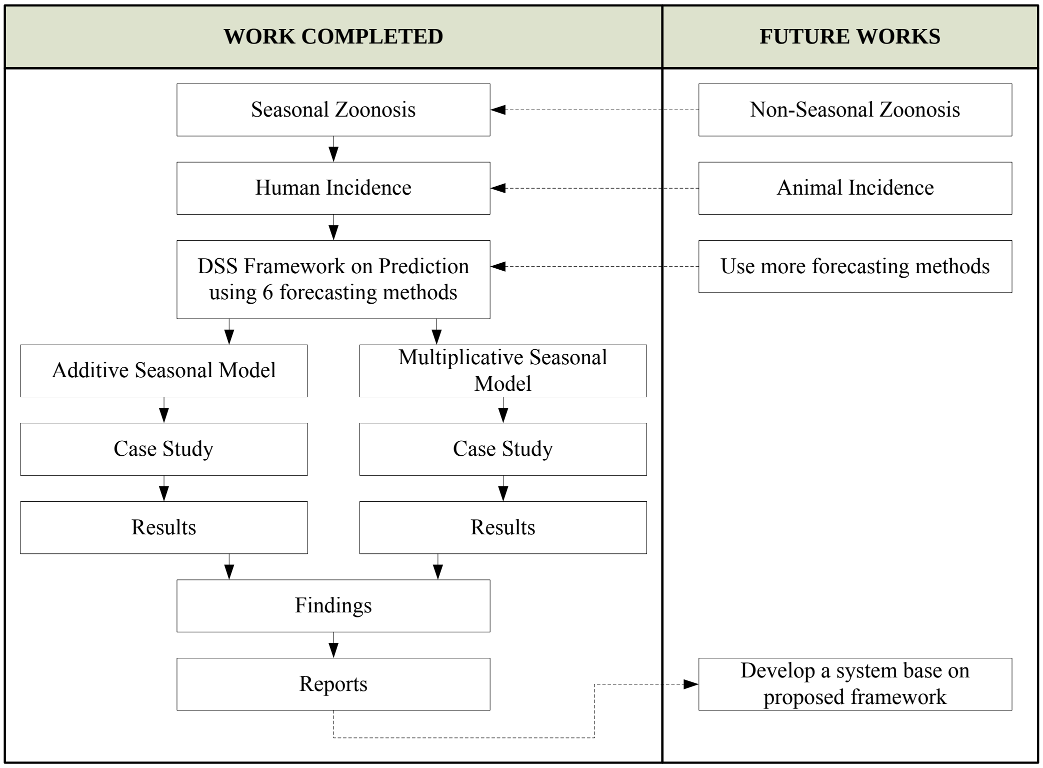


Figure 7.1 Flow Diagrams of Work Completed and Future Works

REFERENCES

[1] CDC, "Compendium of Measures To Prevent Disease Associated with Animals in Public Settings," National Association of State Public Health Veterinarians, Inc. (NASPHV) MMWR 2005;54 (No. RR-4), 2005.

[2] B. A. Wilcox and R. R. Colwell, "Emerging and Reemerging Infectious Diseases: Biocomplexity as an Interdisciplinary Paradigm," *EcoHealth*, vol. 2, pp. 244-257, 2005.

[3] S. Palmer, D. Brown, and D. Morgan, "Early qualitative risk assessment of the emerging zoonotic potential of animal diseases," *BMJ*, vol. 331, pp. 1256-1260, 2005.

[4] C. Brown, "Emerging zoonoses and pathogens of public health significance - an overview," *Rev. sci. tech. Off. int. Epiz.*, vol. 23, pp. 435-442, 2004.

[5] T. Murray, "New U.K. centre takes aim at zoonotic illnesses," in *Medical Post*, vol. 42. Toronto, 2006, pp. 51.

[6] WHO, "Report of the WHO/FAO/OIE joint consultation on emerging zoonotic diseases," Geneva, Switzerland, 3–5 May 2004.

[7] J. Slingenbergh, M. Gilbert, K. de Balogh, and W. Wint, "Ecological sources of zoonotic diseases," *Rev. sci. tech. Off. int. Epiz.*, vol. 23, pp. 467-484, 2004.

[8] WHO. (2007). Zoonoses and veterinary public health (VPH) [Online]. Available: http://www.who.int

[9] WHO. (2010). Pandemic (H1N1) 2009 - update 99 [Online]. Available: http://www.who.int/csr/don/2010_05_07/en/index.html

[10] WHO. (2009). Disease Outbreak News [Online]. Available: http://www.who.int/csr/don/en/

[11] B. A. Wilcox and D. J. Gubler, "Disease Ecology and the Global Emergence of Zoonotic Pathogens," *Environmental Health and Preventive Medicine*, vol. 10, pp. 263–272, 2005.

[12] C. Stephen, H. Artsob, W. Bowie, and D. Patrick, "Perspectives on emerging zoonotic disease research and capacity building in Canada," *The Canadian Journal of Infectious Diseases and Medical Microbiology*, vol. 15, pp. 339-344, 2004.

[13] D. Zeng, H. Chen, C. Tseng, C. Larson, M. Eidson, I. Gotham, C. Lynch, and M. Ascher, "Sharing and visualizing infectious disease datasets using the WNV-BOT portal system," in *Proc The 2004 Annual national conference on Digital government research*, Seattle, 2004, pp. 1-2.

[14] B. Deal, C. Farello, M. Lancaster, T. Kompare, and B. Hannon, "A dynamic model of the spatial spread of an infectious disease: the case Of Fox Rabies in Illinois," *Environmental Modeling and Assessment*, vol. 5, pp. 47-62, 2000.

[15] D. U. Pfeiffer, R. S. Morris, and R. L. Sanson, "Application of GIS in animal disease control-possibilities and limits," in *Proc WHO Consultation on Development and Application of Geographical Methods in the Epidemiology of Zoonoses*, Wusterhasen, Germany, 1994, pp. 1-14.

[16] S. Palmer, D. Brown, and D. Morgan, "Early qualitative risk assessment of the emerging zoonotic potential of animal diseases," *BMJ*, vol. 331, 2005.

[17] G. K. Brückner, W. Vosloo, B. J. A. Du Plessis, P. E. L. G. Kloeck, L. Connoway, M. D. Ekron, D. B. Weaver, C. J. Dickason, F. J. Schreuder, T. Marais, and M. E. Mogajane, "Foot and mouth disease: the experience of South Africa," *Rev. sci. tech. Off. int. Epiz.*, vol. 21, pp. 751-764, 2002.

[18] E. Symeonakis, T. Robinson, and N. Drake, "GIS and multiple-criteria evaluation for the optimisation of tsetse fly eradication programmes," *Environ Monit Assess*, vol. 124, pp. 89-103, 2006.

[19] D. U. Pfeiffer and M. Hugh-Jones, "Geographical information system as a tool in epidemiological assessment and wildlife disease management," *Rev. sci. tech. Off. int. Epiz.*, vol. 21, pp. 91-102, 2002.

[20] V. W. Lees, "Learning from outbreaks of bovine tuberculosis near Riding Mountain National Park: Applications to a foreign animal disease outbreak," *Can Vet J*, vol. 45, pp. 28-34, 2004.

[21] N. Taylor, "Review of the use of models in informing disease control policy development and adjustment," DEFRA, U.K. 26 May 2003.

[22] R. G. Bengis, F. A. Leighton, J. R. Fischer, M. Artois, T. Mörner, and C. M. Tate, "The role of wildlife in emerging and re-emerging zoonoses," in *Rev. sci. tech. Off. int. Epiz.*, vol. 23, 2004, pp. 497-511.

[23] WHO, FAO, and OIE, "Report of the WHO/FAO/OIE joint consultation on emerging zoonotic diseases," Geneva, Switzerland, 3–5 May 2004.

[24] E. Etter, P. Donado, F. Jori, A. Caron, F. Goutard, and F. Roger, "Risk Analysis and Bovine Tuberculosis, a Re-emerging Zoonosis," *Ann NY Acad Sci*, vol. 1081, pp. 61-73, 2006.

[25] D. Zeng, H. Chen, C. Tseng, C. A. Larson, M. Eidson, I. Gotham, C. Lynch, and M. Ascher, "Towards A National Infectious Disease Information Infrastructure: A Case Study in West Nile Virus and Botulism," in *Proc The 2004 Annual national conference on Digital government research*, Seattle, 2004, pp. 1-10.

[26] D. M. Bravata, K. M. McDonald, W. M. Smith, C. Rydzak, H. Szeto, D. L. Buckeridge, C. Haberland, and D. K. Owens, "Systematic Review: Surveillance Systems for Early Detection of Bioterrorism-Related Diseases," *Ann Intern Med*, vol. 140, pp. 910-922, 2004.

[27] G. C. Matibag, M. Igarashi, R. E. La Porte, and H. Tamashiro, "Advocacy, promotion and e-learning: supercourse for zoonosis," *Environmental Health and Preventive Medicine*, vol. 10, pp. 273-281, 2005.

[28] S. M. Debanne, R. A. Bielefeld, G. M. Cauthen, T. M. Daniel, and D. Y. Rowland, "Multivariate Markovian Modeling of Tuberculosis: Forecast for the United States," *Emerging Infectious Diseases*, vol. 6, pp. 148-157, 2000.

[29] D. C. Medina, S. E. Findley, and S. Doumbia, "State–Space Forecasting of Schistosoma haematobium Time-Series in Niono, Mali," *PLoS Neglected Tropical Diseases*, vol. 2, pp. 1-12, 2008.

[30] D. Lai, "Monitoring the SARS Epidemic in China: A Time Series Analysis," *Journal of Data Science*, vol. 3, pp. 279-293, 2005.

[31] P. Sebastiani, K. D. Mandl, P. Szolovits, I. S. Kohane, and M. F. Ramoni, "A Bayesian Dynamic Model for Influenza Surveillance," *Statistics in Medicine*, vol. 25, pp. 1803-1825, 2006.

[32] L. F. Chaves and M. Pascual, "Climate Cycles and Forecasts of Cutaneous Leishmaniasis, a Nonstationary Vector-Borne Disease," *PLoS Medicine*, vol. 3, pp. 1320-1328, 2006.

[33] P. Shankar, M. Walji, A. Ali, and K. A. Johnson-Throop, "Decision support systems to identify different species of malarial parasites," in *Proc AMIA 2003*, 2003, pp. 1006.

[34] X. Y. Sai, Z. Y. Zhang, D. Z. Xu, Y. P. Yan, L. S. Li, and X. N. Zhou, "Application of "time series analysis" in the prediction of schistosomiasis prevalence in areas of "breaking dikes or opening sluice for waterstore" in Dongting Lake areas, China," *Zhonghua Liu Xing Bing Xue Za Zhi.* , vol. 25, pp. 863-866, 2004.

[35] O. J. Briët, P. Vounatsou, D. M. Gunawardena, G. N. Galappaththy, and P. H. Amerasinghe, "Models for short term malaria prediction in Sri Lanka," *Malaria Journal*, vol. 7, pp. doi: 10.1186/1475-2875-7-76, 2008.

[36] WHO. (2009). Global Early Warning System for Major Animal Diseases, including Zoonoses (GLEWS) [Online]. Available: http://www.who.int/zoonoses/outbreaks/glews/en/

[37] M. J. Keeling, "Models of foot-and-mouth disease," *The Royal Society*, vol. 272, pp. 1195–1202, 2005.

[38] J. M. Scudamore and D. M. Harris, "Control of foot and mouth disease: lessons from the experience of the outbreak in Great Britain in 2001," *Rev Sci Tech*, vol. 21, pp. 699-710, 2002.

[39] M. Koopmans, B. Wilbrink, M. Conyn, G. Natrop, H. van der Nat, H. Vennema, A. Meijer, J. van Steenbergen, R. Fouchier, A. Osterhaus, and A. Bosman, "Transmission of H7N7 avian influenza A virus to human beings during a large outbreak in commercial poultry farms in the Netherlands," *The Lancet*, vol. 363, pp. 587-593, 2004.

[40] L. Jinping, R. Qianlu, C. Xi, and Y. Jianqin, "Study on transmission model of avian influenza," in *Proc International Conference on Information Acquisition 2004*, China, 2004, pp. 54-58.

[41] R. M. Bush, "Influenza as a model system for studying the cross-species transfer and evolution of the SARS coronavirus," *The Royal Society*, vol. 359, pp. 1067–1073, 2004.

[42] J. Arikawa, K. Yoshimatsu, U. T. Truong, and U. N. Truong, "Hantavirus Infection - typical rodent-borne viral zoonosis," *Tropical Medicine and Health*, vol. 35, pp. 55-59, 2007.

[43] E. C. Zielinski-Gutierrez and M. H. Hayden, "A Model for Defining West Nile Virus Risk Perception Based on Ecology and Proximity " *EcoHealth*, vol. 3, pp. 28-34, 2006.

[44] M. G. Garner, G. D. Hess, and X. Yang, "An integrated modelling approach to assess the risk of wind-borne spread of foot-and-mouth disease virus from infected premises," *Environmental Modelling and Assessment*, vol. 11, pp. 195-207, 2005.

[45] A. Hailu, A. Mudawi Musa, C. Royce, and M. Wasunna, "Visceral Leishmaniasis: New Health Tools Are Needed," *PLoS Medicine*, vol. 2, pp. 590-594, 2005.

[46] A. Croft and R. Archer, "Dog Bites in Bosnia," *British Journal of General Practice*, vol. 47, pp. 435-437, 1997.

[47] I. N. Kandun, "Draft Summary Report of Investigation of a suspected unusual event of Dying Crows of unknown Etiology in post tsunami, Maldives, 2005," STP-CSR WHO/SEARO, Delhi 2005.

[48] Soeharsono, *Zoonosis Penyakit Menular dari Hewan ke Manusia*, vol. 2. Yogyakarta: Kanisius, 2005.

[49] Soedarto, *Zoonosis Kedokteran*. Surabaya: Airlangga University Press, 2003.

[50] V. P. Health. (2008). Detection, surveillance and control of zoonoses and other relevant diseases [Online]. Available: http://www.veterinary-public-health.de/home_e/aufgaben/zoonosen/zoonosen_e.htm

[51] D. D. Kapan, S. N. Bennett, B. N. Ellis, J. Fox, N. D. Lewis, J. H. Spencer, S. Saksena, and B. A. Wilcox, "Avian Influenza (H5N1) and the Evolutionary and Social Ecology of Infectious Disease Emergence," *EcoHealth*, vol. 3, pp. 187-194, 2006.

[52] D. Despommier, B. R. Ellis, and B. A. Wilcox, "The Role of Ecotones in Emerging Infectious Diseases," *EcoHealth*, vol. 3, pp. 281-289, 2006.

[53] J. N. S. Eisenberg, M. A. Desai, K. Levy, S. J. Bates, S. Liang, K. Naumoff, and J. C. Scott, "Environmental Determinants of Infectious Disease: A Framework for Tracking Causal Links and Guiding Public Health Research," *Environmental Health Perspectives*, vol. 115, pp. 1216-1223, 2007.

[54] J. A. Patz, D. Campbell-Lendrum, T. Holloway, and J. A. Foley, "Impact of regional climate change on human health," *Nature*, vol. 438, pp. 310-317, 2005.

[55] T. N. Palmer, F. J. Doblas-Reyes, R. Hagedorn, and A. Weisheimer, "Probabilistic prediction of climate using multi-model ensembles: from basics to applications," *Philosophical Transaction of the Royal Society B*, vol. 360, pp. 1991-1998, 2005.

[56] L. Stone, R. Olinky, and A. Huppert, "Seasonal dynamics of recurrent epidemics," *Nature*, vol. 446, pp. 533-536, 2007.

[57] L. J. King, N. Marano, and J. M. Hughes, "New partnerships between animal health services and public health agencies," *Rev. sci. tech. Off. int. Epiz.*, vol. 23, pp. 717-725, 2004.

[58] G. Enteric Zoonotic Disease Modelling, "A research plan for enteric zoonoses: Modelling the link between human health and the environment; identifying effective Interventions," August 2006.

[59] E. S. Shtatland and T. Shtatland, "Why We Need a Bayesian Approach to Early Detection of Epidemic Outbreaks and Financial Bubbles Using First-Order Autoregressive Models with Structural Changes," in *Proc NESUG 2009*, Burlington, Vermont, 2009, pp. 1-12.

[60] J. K. Taubenberger and D. M. Morens, "1918 Influenza: the mother of all pandemics," *Emerging Infectious Diseases*, vol. 12, pp. 15-22, 2006.

[61] J. C. Gaydos, F. H. Top, R. A. Hodder, and P. K. Russell, "Swine influenza a outbreak, Fort Dix, New Jersey, 1976," *Emerging Infectious Diseases*, vol. 12, pp. 23-28, 2006.

[62] K. Kimura, A. Adlakha, and P. M. Simon, "Fatal case of swine influenza virus in an immunocompetent host," *Mayo Clinic Proceedings. Mayo Clinic*, vol. 73, pp. 243-245, 1998.

[63] CDC, "Swine Influenza A (H1N1) Infection in Two Children --- Southern California, March--April 2009," 24 April 2009.

[64] CDC, "Outbreak of Swine-Origin Influenza A (H1N1) Virus Infection --- Mexico, March--April 2009," 30 April 2009.

[65] WHO. (2009). Pandemic (H1N1) 2009 - update 76 [Online]. Available: http://www.who.int/csr/don/2009_11_27a/ar/index.html

[66] WHO. (2004). Avian influenza ("bird flu") and the significance of its transmission to humans [Online]. Available: http://www.who.int/csr/don/2004_01_15/en/

[67] WHO. (2010). Cumulative Number of Confirmed Human Cases of Avian Influenza A/(H5N1) Reported to WHO [Online]. Available: http://www.who.int/csr/disease/avian_influenza/country/cases_table_2010_05_06/en/index.html

[68] WHO. (2010). Drug-resistant Salmonella [Online]. Available: http://www.who.int/mediacentre/factsheets/fs139/en/

[69] WHO. (2010). Tuberculosis-Fact Sheet [Online]. Available: http://www.who.int/mediacentre/factsheets/fs104/en/index.html

[70] J. Mackenzie, "Emerging Viral Diseases in South-East Asia and the Western Pacific: the Importance of Biosecurity and the Dilemma of Dual-Use," in *Proc Singapore Workshop 2007*, Singapore, 2007, pp.

[71] M. Helms, P. Vastrup, P. Gerner-Smidt, and K. Mølbak, "Short and long term mortality associated with foodborne bacterial gastrointestinal infections: registry based study," *British Medical Journal*, vol. 326, pp. 357-361, 2003.

[72] W. G. Scott, H. M. Scott, R. J. Lake, and M. Baker, "Economic cost to New Zealand of foodborne infectious disease," *New Zealand Medical Journal*, vol. 113, pp. 281-284, 2000.

[73] J. M. Drake, "Limits to Forecasting Precision for Outbreaks of Directly Transmitted Diseases," *PLoS Medicine*, vol. 3, pp. 57-62, 2005.

[74] E. S. Shtatland, "Low-Order Autoregressive Models in Early Detection of Epidemic Outbreaks and Explosive Behaviors in Economic and Financial Time Series," in *Proc NESUG 2007*, Baltimore, Maryland, 2007, pp. 1-10.

[75] E. S. Shtatland, K. Kleinman, and E. M. Cain, "Biosurveillance and outbreak detection using the ARIMA and logistic procedures," in *Proc SUGI 31*, Cary, NC: SAS Institute, Inc, 2006, pp. 197-31.

[76] P. M. Luz, B. V. M. Mendes, C. T. Codeço, C. J. Struchiner, and A. P. Galvani, "Time Series Analysis of Dengue Incidence in Rio de Janeiro, Brazil," *American Journal of Tropical Medicine and Hygiene*, vol. 79, pp. 933-939, 2008.

[77] A. Earnest, M. I. Chen, D. Ng, and L. Y. Sin, "Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore," *BMC Health Services Research*, vol. 5, pp. doi:10.1186/1472-6963-5-36, 2005.

[78] A. Gomez-Elipe, A. Otero, M. van Herp, and A. Aguirre-Jaime, "Forecasting malaria incidence based on monthly case reports and environmental factors in Karuzi, Burundi, 1997-2003," *Malaria Journal*, vol. 6, pp. doi:10.1186/1475-2875-6-129, 2007.

[79] T. A. Abeku, S. J. d. Vlas, G. Borsboom, A. Teklehaimanot, A. Kebede, D. Olana, G. J. van Oortmarssen, and J. D. F. Habbema, "Forecasting malaria incidence from historical morbidity patterns in epidemic-prone areas of Ethiopia: a simple seasonal adjustment method performs best " *Tropical Medicine and International Health*, vol. 7, pp. 851-857, 2002.

[80] S. Thammapalo, V. Chongsuwiwatwong, D. McNeil, and A. Geater, "The Climatic Factors Influencing The Occurrence of Dengue Hemorrhagic Fever in Thailand," *Southeast Asian journal of tropical medicine and public health*, vol. 36, 2005.

[81] H. D. Teklehaimanot, J. Schwartz, A. Teklehaimanot, and M. Lipsitch, "Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia II. Weather-based prediction systems perform comparably to early detection systems in identifying times for interventions," *Malaria Journal*, vol. 3, pp. 10.1186/1475-2875-3-44, 2004.

[82] S. Konchom, P. Singhasivanon, J. Kaewkungwa, S. Chuprapawan, K. Thimasarn, C. Kidson, S. Yimsamran, and C. Rojanawatsirivet, "Early Detection of Malaria in an Endemic Area: Model Development," *Southeast Asian journal of tropical medicine and public health*, vol. 37, pp. 1067-1071, 2006.

[83] K. K. Chui, P. Webb, R. M. Russell, and E. N. Naumova, "Geographic variations and temporal trends of Salmonella-associated hospitalization in the U.S. elderly, 1991-2004: A time series analysis of the impact of HACCP regulation," *BMC Public Health*, vol. 9, pp. doi: 10.1186/1471-2458-9-447, 2009.

[84] L. F. Chaves and M. Pascual, "Comparing Models for Early Warning Systems of Neglected Tropical Diseases," *PLoS Neglected Tropical Diseases*, vol. 1, pp. e33, 2007.

[85] T. A. Hammad, M. F. Abdel-Wahab, N. DeClaris, A. El-Sahly, and N. El-Kady, "Comparative evaluation of the use of artificial neural networks for modelling

the epidemiology of schistosomiasis mansoni," *Trans R Soc Trop Med Hyg.*, vol. 90, pp. 372-6, 1996.

[86] C. Daniel, Medina, S. E. Findley, B. Guindo, and S. Doumbia, "Forecasting Non-Stationary Diarrhea, Acute Respiratory Infection, and Malaria Time-Series in Niono, Mali," *Plos One*, vol. 2, pp. e1181, 2007.

[87] R. Maijala and J. Ranta, "The Use of Predictive Models to Manage Risks Caused by Salmonella in Broilers," in *Proc Fosare Seminar Series 1 on Newly Emerging Pathogens, Including Risk Assessment And Risk Management*, Brussels, 2003, pp. 13-15.

[88] L. Tanner, M. Schreiber, J. G. H. Low, A. Ong, T. Tolfvenstam, Y. L. Lai, L. C. Ng, Y. S. Leo, L. T. Puong, S. G. Vasudevan, C. P. Simmons, M. L. Hibberd, and E. E. Ooi, "Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness," *PLoS Neglected Tropical Diseases*, vol. 2, pp. e196, 2008.

[89] Medical Dictionary. (2004). Definition of Disease surveillance [Online]. Available: http://www.medterms.com/script/main/art.asp?articlekey=26702

[90] UNC. (2003). Spatial-Temporal Model [Online]. Available: www.stat.unc.edu/faculty/rs/s321/spatemp.pdf

[91] C. Beck-Wörner, G. Raso, P. Vounatsou, E. K. N'goran, G. Rigo, E. Parlow, and J. Utzinger, "Bayesian Spatial Risk Prediction of Schistosoma Mansoni Infection In Western Côte D'ivoire Using a Remotely-Sensed Digital Elevation Model," *American Journal of Tropical Medicine and Hygiene*, vol. 76, pp. 956-963, 2007.

[92] G. J. Yang, A. Gemperli, P. Vounatsou, M. Tanner, X.-N. Zhou, and J. Utzinger, "A Growing Degree-Days Based Time-Series Analysis for Prediction of Schistosoma Japonicum Transmission In Jiangsu Province, China," *American Journal of Tropical Medicine and Hygiene*, vol. 75, pp. 549-555, 2006.

[93] F. Rakotomanana, R. V. Randremanana, L. P. Rabarijaona, J. B. Duchemin, J. Ratovonjato, F. Ariey, J. P. Rudant, and I. Jeanne, "Determining areas that require indoor insecticide spraying using Multi Criteria Evaluation, a decision-support tool for malaria vector control programmes in the Central Highlands of Madagascar," *International Journal of Health Geographics*, vol. 6, pp. doi:10.1186/1476-072X-6-2, 2007.

[94] L. Li, L. Bian, and G. Yun, "A study of the distribution and abundance of the adult malaria vector in western Kenya highlands," *International Journal of Health Geographics*, vol. 7, pp. doi:10.1186/1476-072X-7-50, 2008.

[95] A. Daash, A. Srivastava, B. N. Nagpal, R. Saxena, and S. K. Gupta, "Geographical information system (GIS) in decision support to control malaria – a case study of Koraput district in Orissa, India," *Journal Vector Borne Disease*, vol. 46, pp. 72-74, 2009.

[96] E. N. Naumova, E. O'Neil, and I. MacNeill, "INFERNO: A System for Early Outbreak Detection and Signature Forecasting," *MMWR Supplement* vol. 54, pp. 77-83, 2005.

[97] D. J. Power. (2007). A Brief History of Decision Support Systems [Online]. Available: http://dssresources.com/history/dsshistory.html

[98] E. Turban, J. E. Aronson, and T. P. Liang, *Decision Support System and Intelligent Systems*, 7th ed. New Jersey: Prentice Hall, 2005.

[99] G. M. Marakas, *Decision Support Systems in The 21st Century*. New Jersey: Prentice Hall, 1999.

[100] C. W. Holsapple and A. B. Whinston, *Decision Support Systems: A Knowledge Based Approach*, 10th ed: West Group, 1996.

[101] A. P. Sage, *Decision Support System Engineering*. New York, 1991.

[102] R. H. Sprague, Jr and E. D. Carlson, *Building Effective Decision Support Systems*: NJ: Prentice Hall, 1982.

[103] E. Turban and J. E. Aronson, "A schematic view of DSS ", A. s. v. o. DSS, Ed., 2000.

[104] B. L. Bowerman and R. T. O'Connell, *Forecasting and Time Series An Applied Approach*, 3rd ed: Duxbury Thomson Learning, 1993.

[105] ChartSchool. (2010). Moving Averages [Online]. Available: http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:moving_averages

[106] P. A. Jensen. (2004). Forecasting Theory [Online]. Available: http://www.me.utexas.edu/~jensen/ORMM/omie/operation/unit/forecast/index.html

[107] S. Makridakis and S. C. Wheelwright, *Forecasting Methods and Applications*: John Wiley & Sons. Inc, 1978.

[108] P. S. Kalekar. (2004). Time series Forecasting using Holt-Winters Exponential Smoothing [Online]. Available: http://www.it.iitb.ac.in/~praj/acads/seminar/04329008_ExponentialSmoothing.pdf

[109] StatSoft. (2009). Time Series Analysis [Online]. Available: http://www.statsoft.com/textbook/time-series-analysis/

[110] S. Dave. (2004). Forecasting [Online]. Available: http://www.uoguelph.ca/~dsparlin/forecast.htm

[111] H. Tim, O. C. Marcus, and R. William, "Neural Network Models for Time Series Forecasts," *Management Science*, vol. 42, pp. 1082-1092, 1996.

[112] Investors Intelligence. (2010). Moving average [Online]. Available: http://www.chartanalysts.com/x/moving_averages.html

[113] NIST Sematech. (2006). Engineering Statistics Handbook [Online]. Available: http://www.itl.nist.gov/div898/handbook/toolaids/pff/ehb-chapters-1-8.pdf

[114] D. Tikunov and T. Nishimura, "Traffic prediction for mobile network using Holt-Winter's exponential smoothing," in *Proc Software, Telecommunications and Computer Networks, 2007. SoftCOM 2007. 15th International Conference on*, 2007, pp. 1-5.

[115] J. G. Caldwell. (2006). The Box-Jenkins Forecasting Technique [Online]. Available: http://www.foundationwebsite.org/

[116] C. Chia-Lin, S. Songsak, and W. Aree, "Modelling and forecasting tourism from East Asia to Thailand under temporal and spatial aggregation," *Math. Comput. Simul.*, vol. 79, pp. 1730-1744, 2009.

[117] L. Quantitative Micro Software. (2005). EViews 5.1 User's Guide [Online]. Available: www.eviews.com

[118] C. Frank, A. Garg, L. Sztandera, and A. Raheja, "Forecasting women's apparel sales using mathematical modeling," *International Journal of Clothing Science and Technology*, vol. 15, pp. 107-125, 2003.

[119] Y. Jin. (2008). A Definition of Soft Computing - adapted from L.A. Zadeh [Online]. Available: http://www.soft-computing.de/def.html

[120] S. Samarasinghe, *Neural Networks for Applied Sciences and Engineering*. USA: Taylor and Francis Group, 2007.

[121] G. P. Zang and M. Qi, "Neural network forecasting for seasonal and trend time series," *European Journal of Operational Research*, vol. 160, pp. 501-514, 2005.

[122] L. O. Teles, V. Vasconcelos, E. Pereira, and M. Saker, "Time Series Forecasting of Cyanobacteria Blooms in the Crestuma Reservoir (Douro River, Portugal) Using Artificial Neural Network," *Environmental Management*, vol. 38, pp. 227-237, 2006.

[123] A. F. Atiya, S. M. El-Shoura, S. I. Shaheen, and M. S. El-Sherif, "A Comparison Between Neural-Network Forecasting Techniques—Case Study: River Flow Forecasting," *IEEE Transactions on Neural Networks*, vol. 10, pp. 402-409, 1999.

[124] StatSoft. (2008). ANOVA/MANOVA [Online]. Available: http://www.statsoft.com/textbook/stanman.html

[125] A. Agresti. (2007). Chapter 12 Comparing Groups: Analysis of Variance (ANOVA) Methods [Online]. Available: www.stat.ufl.edu/~aa/sta6127/ch12.pdf

[126] V. Bewick, L. Cheek, and J. Ball, "Statistics review 9: One-way analysis of variance," *Critical Care*, vol. 8, pp. 130-136, 2004.

[127] J. P. Geaghan. (2009). [Online]. Available: http://www.stat.lsu.edu/faculty/geaghan/EXST7015/Fall2009/PDF/Lecture19a%20Notes%20Fall2009.pdf

[128] GraphPad Software Inc. (2007). Coefficient of variation (CV) [Online]. Available: http://www.graphpad.com/help/prism5/prism5help.html?coefficient_of_variation_%28cv%29.htm

[129] M. Gilliland. (2009). The Coefficient of Variation for assessing forecastability [Online]. Available: http://blogs.sas.com/forecasting/index.php?/archives/10-The-Coefficient-of-Variation-for-assessing-forecastability.html

[130] M. J. Druzdzel and R. R. Flynn. (2002). **Decision Support Systems** [Online]. Available: www.pitt.edu/~druzdzel/psfiles/dss.ps.Z

[131] S. C. Albright, *VBA for Modelers: Developing Decision Support Systems with Microsoft Excel*: Duxburry, Thomson Learning, 2001.

[132] J. Walkenbach, "Excel VBA Programming for Dummies," Wiley Publishig, Inc., 2004.

[133] WHO. (2009). Salmonella [Online]. Available: http://www.who.int/topics/salmonella/en/

[134] Centers for Disease Control and Prevention, "Compendium of Measures To Prevent Disease Associated with Animals in Public Settings," National Association of State Public Health Veterinarians, Inc. (NASPHV) MMWR 2005;54 (No. RR-4), 2005.

[135] National Institute of Allergy and Infectious Diseases. (2009). Salmonellosis [Online]. Available: http://www3.niaid.nih.gov/topics/salmonellosis/

[136] Intervet/Schering-Plough Animal Health. (2009). Salmonella epidemiology [Online]. Available: http://www.safe-poultry.com/salmonellaepidemiology.asp

[137] Intervet/Schering-Plough Animal Health. (2009). Salmonella in Asia [Online]. Available: http://www.safe-poultry.com/SalmonellainAsia.asp

[138] Intervet/Schering-Plough Animal Health. (2009). Salmonella in the European Union [Online]. Available: http://www.safe-poultry.com/SalmonellaintheEU.asp

[139] Intervet/Schering-Plough Animal Health. (2009). Salmonella in Oceania [Online]. Available: http://www.safe-poultry.com/SalmonellainOceania.asp

[140] WHO. (2007). Tuberculosis-Fact Sheet [Online]. Available: http://www.who.int/mediacentre/factsheets/fs104/en/index.html

[141] O. Cosivi, J. M. Grange, C. J. Daborn, M. C. Raviglione, T. Fujikura, D. Cousins, R. A. Robinson, H. F. A. K. Huchzermeyer, I. d. Kantor, and F.-X. Meslin, "Zoonotic Tuberculosis due to Mycobacterium bovis in Developing Countries," *Emerging Infectious Diseases*, vol. 4, 1998.

[142] National Institute of Allergy and Infectious Diseases. (2008). Tuberculosis (TB) [Online]. Available: http://www3.niaid.nih.gov/topics/tuberculosis/Understanding/overview.htm

[143] Centers for Disease Control and Prevention. (2009). Tuberculosis (TB) [Online]. Available: http://wwwnc.cdc.gov/travel/yellowbook/2010/chapter-5/tuberculosis.aspx

[144] Centers for Disease Control and Prevention. (2010). Vision, Mission, Core Values, and Pledge [Online]. Available: http://www.cdc.gov/about/organization/mission.htm

[145] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 1993," *MMWR 1993*, vol. 42 1994.

[146] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 1994," *MMWR 1994*, vol. 43, 1995.

[147] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 1995," *MMWR 1995*, vol. 44 1996.

[148] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 1996," *MMWR 1996*, vol. 45 1997.

[149] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 1997," *MMWR 1997*, vol. 46 1998.

[150] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 1998," *MMWR 1998*, vol. 47 1999.

[151] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 1999," *MMWR 1999*, vol. 48 2000.

[152] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 2000," *MMWR 2000*, vol. 49 2001.

[153] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 2001," *MMWR 2001*, vol. (53), 2002.

[154] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 2002," *MMWR 2002*, vol. 51, 2003.

[155] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 2003," *MMWR 2003*, vol. 52, 2004.

[156] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 2004," *MMWR 2004*, vol. 53 2005.

[157] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 2005," *MMWR 2005*, vol. 54, 2006.

[158] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 2006," *MMWR 2006*, vol. 55 2007.

[159] Centers for Disease Control and Prevention, "Summary of Notifiable Diseases, United States 2007," *MMWR 2006*, vol. 56 2009.

[160] M. Barigozzi. ARIMA estimation: theory and applications [Online]. Available: www.barigozzi.eu/ARIMA.pdf

[161] L. Iñiguez, M. Hilali, D. L. Thomas, and G. Jesry, "Udder measurements and milk production in two awassi sheep genotypes and their crosses," *Journal of Dairy Science*, vol. 92, pp. 4613-4620, 2009.

LIST OF PUBLICATION

Journal & Book Chapter Publications

1. A. E. Permanasari, D. R. Awang Rambli, and P. D. D. Dominic, "Prediction of Zoonosis Incidence in Human using Seasonal Auto Regressive Integrated Moving Average (SARIMA)," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 5, pp. 103-110, 2009.

2. A. E. Permanasari, D. R. Awang Rambli, and P. D. D. Dominic, "A decision support framework for a zoonosis prediction system: case study of Salmonellosis," *International Journal of Medical Engineering and Informatics (IJMEI)*, in printing.

3. A. E. Permanasari, D. R. Awang Rambli, and P. D. D. Dominic, " Performance of univariate forecasting on seasonal diseases: the case of Tuberculosis," *book chapter on Software Tools and Algorithms for Biological Systems*, Springer Book Series in Advances in Experimental Medicine and Biology, (AEMB), in printing.

Conference Publications

1. A.E. Permanasari, D. R. Awang Rambli, and D. D. Dominic, "A Conceptual Framework for Developing a Zoonosis Emerging System," in *Proc. National Postgraduate Conference (NPC) 2008*, Tronoh, Malaysia, 2008.

2. A.E. Permanasari, D. R. Awang Rambli, and D. D. Dominic, "Construction of Zoonosis Domain Relationship as a Preliminary Stage for Developing a Zoonosis Emerging System," in *Proc. International Symposium on Information Technology 2008 (ITSIM '08)*, Kuala Lumpur, pp. 527-534, 2008.

3. A.E. Permanasari, D. R. Awang Rambli, and D. D. Dominic, “Forecasting of Zoonosis Incidence in Human Using Decomposition Method of Seasonal Time Series," in *Proc. National Postgraduate Conference (NPC) 2009*, Tronoh, Malaysia, 2009.

4. A.E. Permanasari, D. R. Awang Rambli, and P. D. D. Dominic, "A Comparative Study of Univariate Forecasting Methods for Predicting Tuberculosis Incidence on Human," in *Proc. Student Conference on Research and Development (SCOReD 2009)*, Kuala Lumpur, Malaysia, pp. 188-191, 2009.

5. A.E. Permanasari, D. R. Awang Rambli, and P. D. D. Dominic, "Forecasting of Salmonellosis Incidence in Human using Artificial Neural Network (ANN)," in *Proc. The 2nd International Conference on Computer and Automation Engineering (ICCAE) Volume 1*, Singapore, pp. 136-139, 2010.

6. A.E. Permanasari, D. R. Awang Rambli, and P. D. D. Dominic, "Decision Support Conceptual Framework for Zoonosis Emerging System," in *Proc. The 2nd International Conference on Computer and Automation Engineering (ICCAE) Volume 2*, Bali, Indonesia, pp. 469-473, 2010.

7. A.E. Permanasari, D. R. Awang Rambli, and P. D. D. Dominic, “Forecasting Method Selection Using ANOVA and Duncan Multiple Range Tests on Time Series Dataset," in *Proc. 4th International Symposium on Information Technology 2010 (ITSIM ’10) Volume 2*, Kuala Lumpur, Malaysia, pp.941-945, 2010.

APPENDIX A

SAMPLES OF RAW DATA

NOTIFIABLE DISEASES — summary of reported cases, by month, United States, 1993

Disease	Total	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Unk.
AIDS*	103,691	7,153	7,222	21,244	6,725	10,081	8,254	7,597	8,524	9,176	5,072	5,797	6,846	-
Amebiasis	2,970	144	198	202	211	218	228	284	313	272	310	245	345	-
Anthrax	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Aseptic meningitis	12,848	519	532	502	478	698	739	1,569	1,891	1,797	1,909	1,005	1,209	-
Botulism, total	97	6	5	4	5	6	6	11	11	8	16	10	9	-
Brucellosis	120	4	6	6	7	8	6	12	8	3	11	7	42	-
Chancroid†	1,399		401			469			229			300		-
Cholera	18	5	3	1	3	2	-	2	-	1	1	-	-	-
Diphtheria	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Encephalitis, primary infections	919	43	49	45	33	53	45	84	93	158	152	72	92	-
Post-infectious	170	7	20	16	14	17	17	13	15	9	10	14	18	-
Gonorrhea†	439,673		103,178			102,890			120,498			113,107		-
Granuloma inguinale†	19		15			4			-			-		-
Haemophilus influenzae	1,419	99	104	149	97	141	100	105	68	78	126	155	197	-
Hansen disease (leprosy)	187	11	9	20	17	27	10	8	18	14	30	11	12	-
Hepatitis A	24,238	1,739	1,718	1,985	1,739	2,104	1,678	2,132	1,661	1,838	2,535	1,916	3,193	-
Hepatitis B	13,361	772	918	1,000	1,038	1,311	999	1,191	994	1,003	1,207	1,036	1,892	-
Hepatitis, non-A, non-B§	4,786	272	354	337	293	383	301	347	340	364	496	384	915	-
Hepatitis, unspecified	627	44	47	52	43	77	34	61	40	55	49	39	86	-
Legionellosis	1,280	110	81	88	77	112	86	104	107	104	155	88	168	-
Leptospirosis	51	1	8	1	1	3	2	4	7	5	3	8	8	-
Lyme disease	8,257	175	310	323	234	433	524	1,474	1,156	845	840	738	1,205	-
Lymphogranuloma venereum†	285		69			73			71			72		-
Malaria	1,411	51	69	74	118	77	98	144	154	124	200	110	192	-
Measles (rubeola)	312	17	41	27	11	28	72	28	24	17	26	9	12	-
Meningococcal infections	2,637	155	211	311	251	273	180	204	109	121	202	160	460	-
Mumps	1,692	101	132	169	144	186	191	120	80	97	138	126	208	-
Murine typhus fever	25	-	-	1	-	3	-	5	-	4	4	4	4	-
Pertussis (whooping cough)	6,586	214	236	202	239	298	329	777	876	988	1,061	469	897	-
Plague	10	-	-	1	-	2	-	1	4	-	2	-	-	-
Poliomyelitis, paralytic¶	3	-	-	-	-	-	-	1	1	-	-	-	1	
Psittacosis	60	3	7	4	3	8	6	5	7	4	1	7	5	-
Rabies, animal	9,377	408	512	649	801	994	724	876	930	899	1,001	713	870	-
Rabies, human	3	-	-	-	-	-	-	-	1	-	-	-	2	-
Rheumatic fever, acute	112	5	16	12	4	10	28	13	7	1	3	4	9	-
Rocky Mountain spotted fever	456	9	3	3	3	16	38	90	95	69	74	24	32	-
Rubella (German measles)	192	10	8	25	15	30	20	30	19	7	5	6	17	-
Rubella, congenital syndrome	5	1	1	-	2	-	-	1	-	-	-	-	-	-
Salmonellosis	41,641	1,909	2,099	2,196	2,188	3,131	3,256	4,819	5,119	4,367	4,980	3,164	4,413	-
Shigellosis	32,198	1,224	1,554	1,507	1,572	2,511	2,799	3,494	3,665	3,060	4,005	2,725	4,082	-
Syphilis, total all stages†	101,259		25,621			26,942			24,692			24,004		-
Primary and secondary†	26,498		6,952			6,684			6,621			6,241		-
Congenital <1 year†	3,211		678			834			864			835		-
Tetanus	48	1	1	1	3	5	5	4	7	5	4	3	9	-
Toxic-shock syndrome	212	7	21	26	23	16	15	19	25	23	15	6	16	-
Trichinosis	16	1	3	3	-	-	1	-	2	-	2	1	3	-
Tuberculosis	25,313	778	1,322	1,881	2,105	1,979	2,371	2,003	2,009	1,938	2,028	1,769	5,130	-
Tularemia	132	3	5	4	3	13	20	24	21	12	12	2	13	-
Typhoid fever	440	26	24	26	25	40	26	46	38	46	62	30	51	-
Varicella (chickenpox)	134,722	12,815	15,322	18,553	17,373	23,933	13,866	5,129	3,093	2,534	4,038	6,536	11,530	-

* AIDS total updated through December 31, 1993.

† Cases updated through Feburary 28, 1994.

§The number of reported cases of non-A, non-B hepatitis is misleading because in some states, reported cases included persons positive for antibody to hepatitis C virus (anti-HCV) identified in routine screening programs but who did not have acute hepatitis.

¶ Subject to change due to retrospective case evaluations or late reports.

APPENDIX B

FORECASTING PLOTS

B1. Forecasting Plots for Salmonellosis

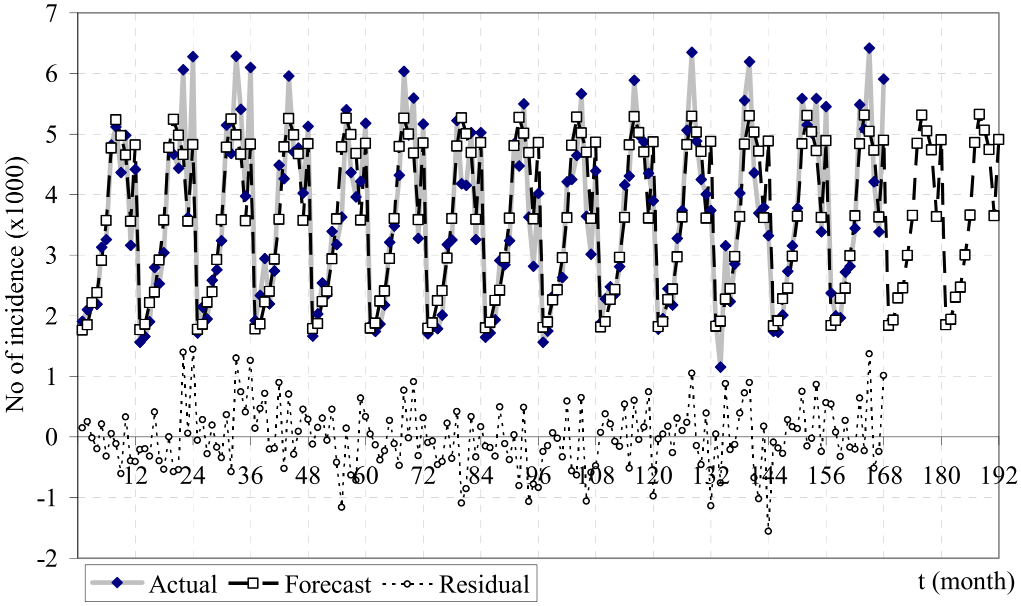


Figure B1.1 Regression Result of Salmonellosis

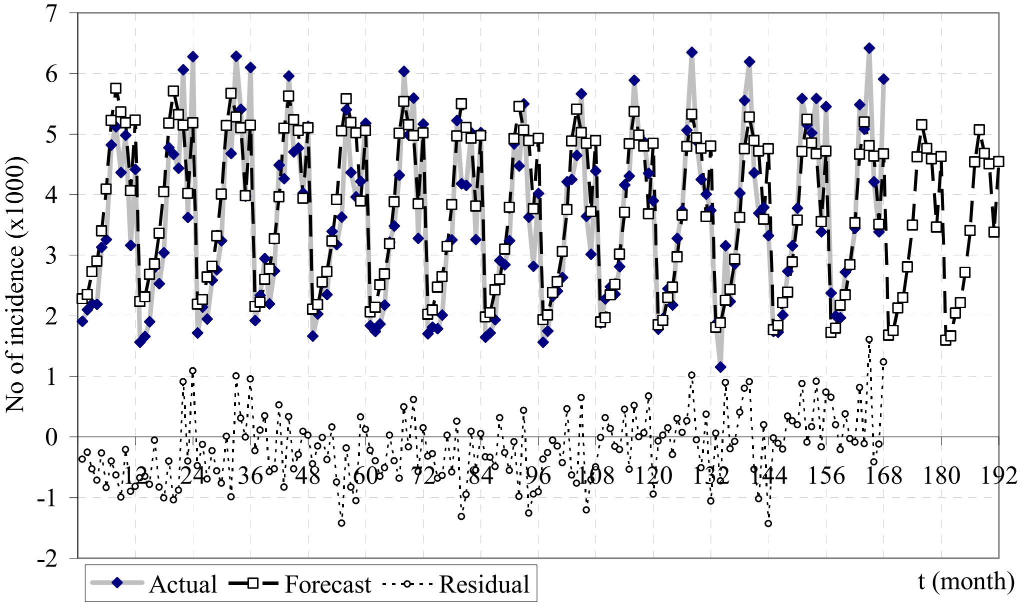


Figure B1.2 Decomposition Result of Salmonellosis

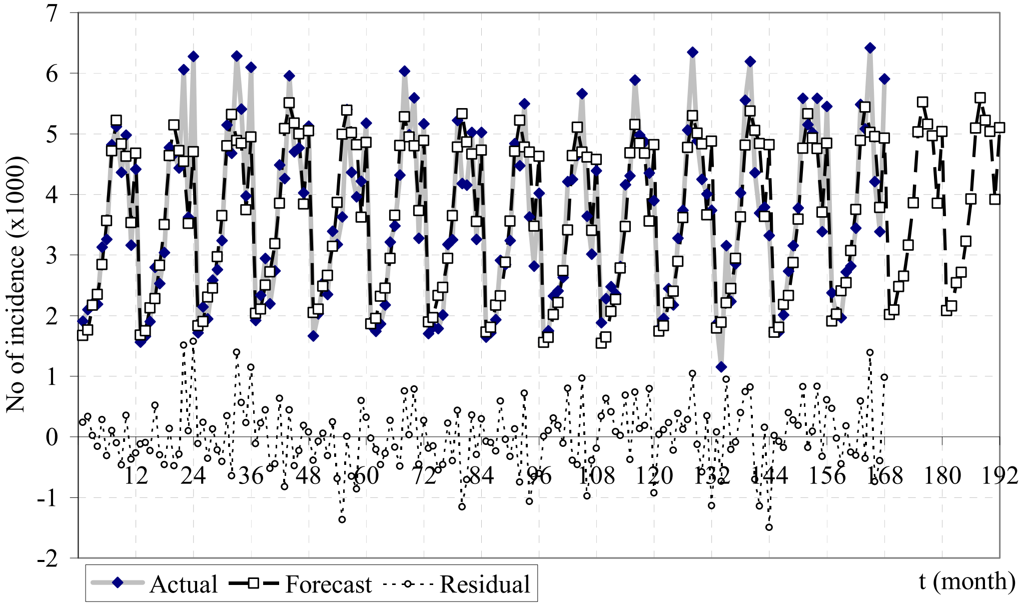


Figure B1.3 Holt-Winter's Result of Salmonellosis

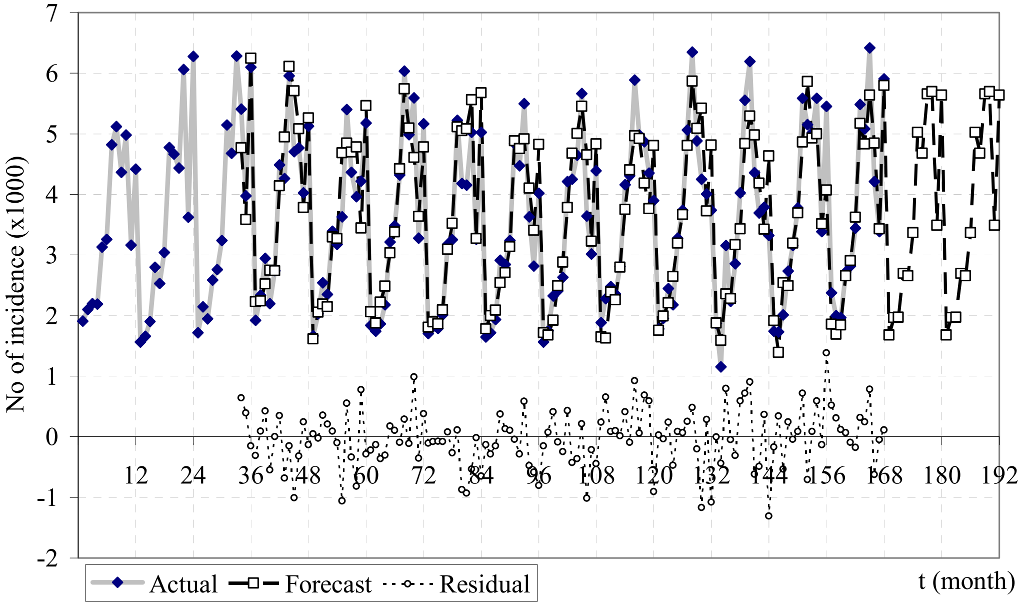


Figure B1.4 ARIMA Result of Salmonellosis

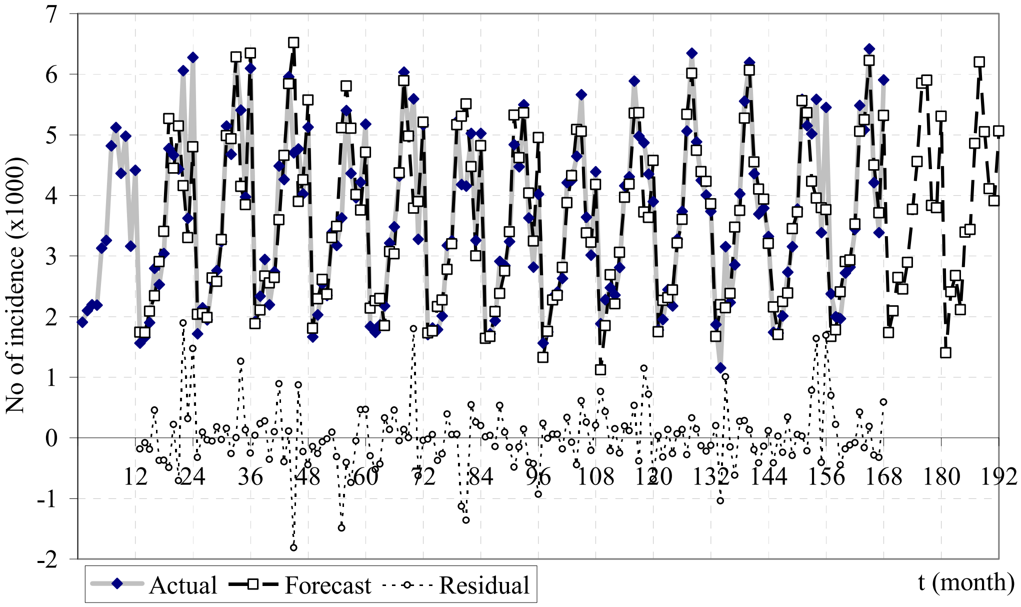


Figure B1.5 Neural Network Result of Salmonellosis

B2. Forecasting Plots for Tuberculosis

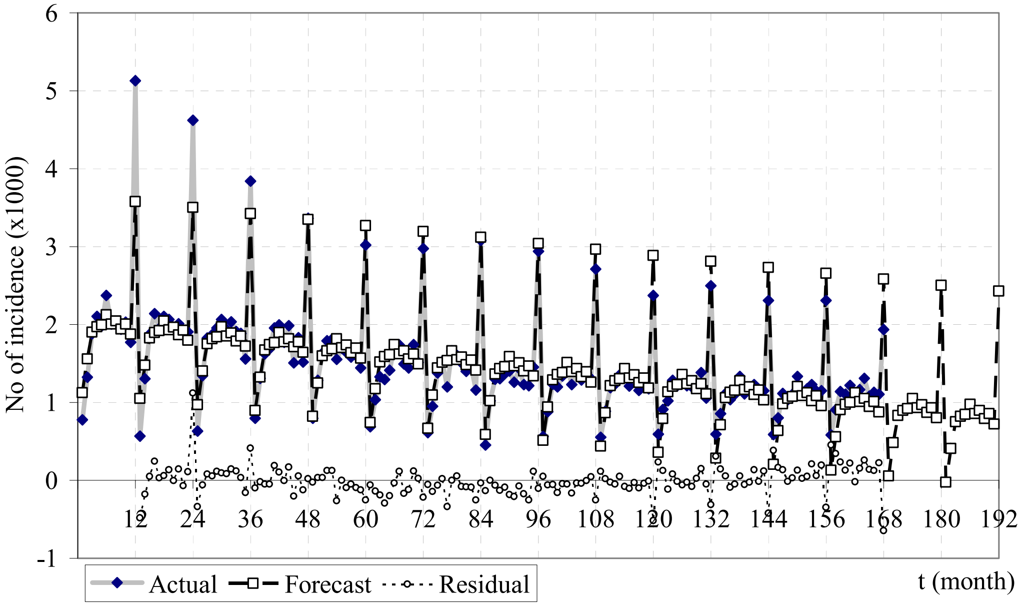


Figure B2.1 Regression Result of Tuberculosis

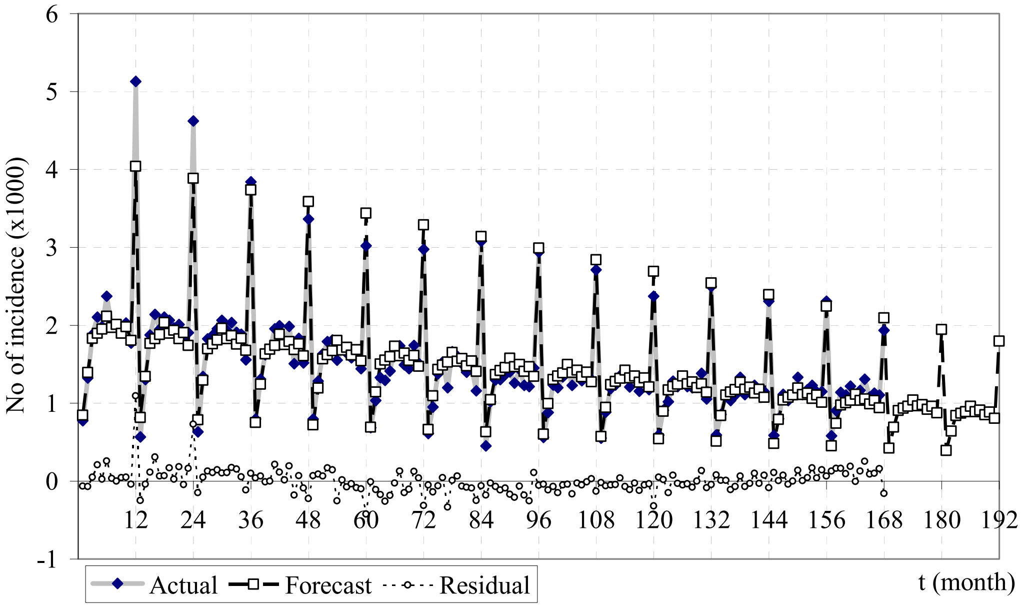


Figure B2.2 Decomposition Result of Tuberculosis

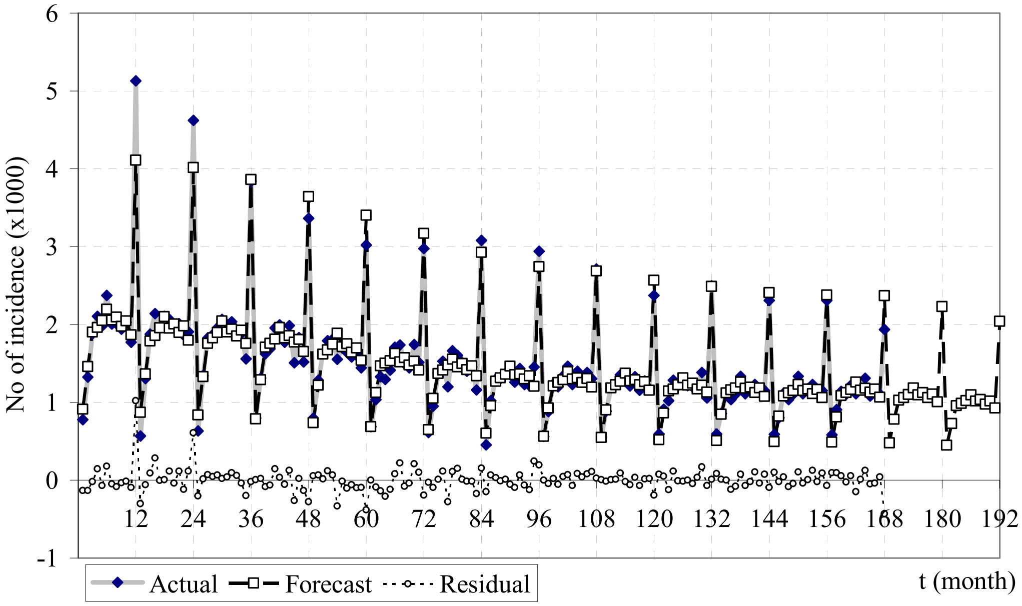


Figure B2.3 Holt-Winter's Result of Tuberculosis

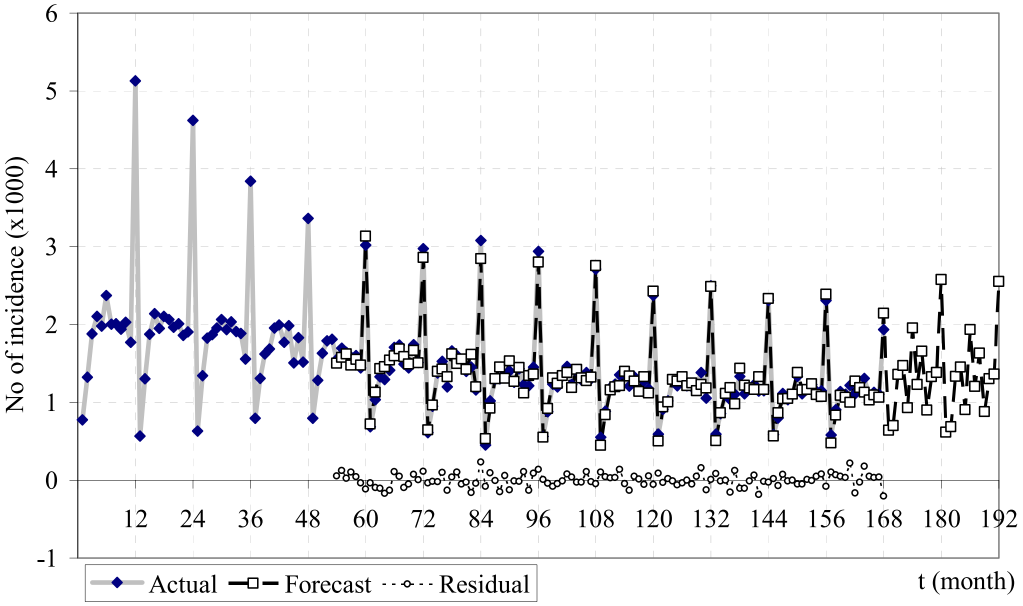


Figure B2.4 ARIMA Result of Tuberculosis

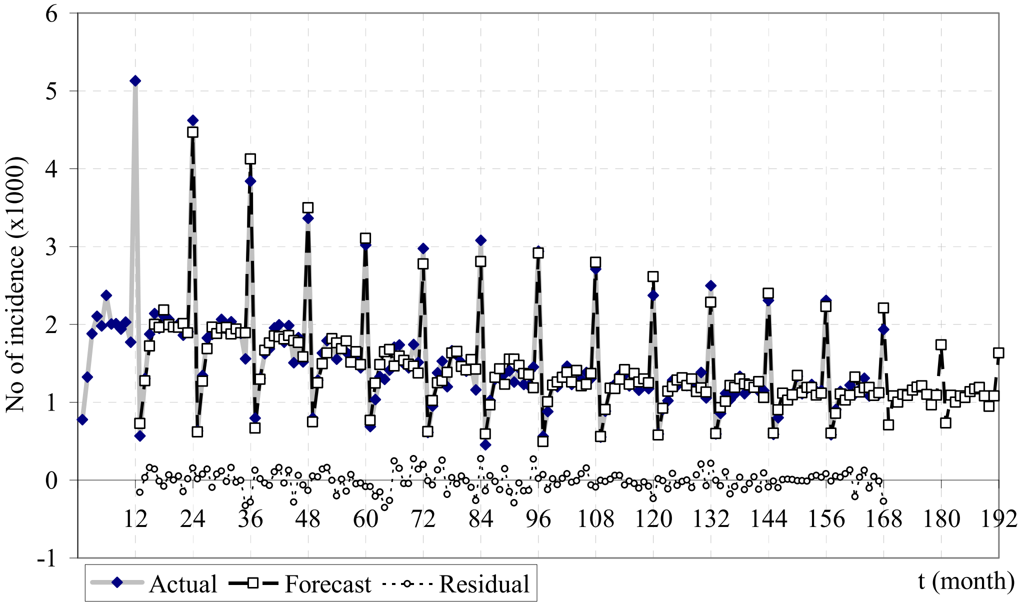


Figure B2.5 Neural Network Result of Tuberculosis

APPENDIX C

TABLE OF SENSITIVITY ANALYSIS

C1. What If (Sensitivity) Analysis Result of Salmonellosis

Table C1.1 Sensitivity of Regression Salmonellosis

Month	Forecast 2008			Forecast 2009		
	Data 93-06	Data 93-07	Senst (%)	Data 93-06	Data 93-07	Senst. (%)
Jan.	1846.680	1813.248	-1.84%	1852.323	1818.891	-1.84%
Feb.	1933.323	1899.891	-1.76%	1938.966	1905.534	-1.75%
Mar.	2297.323	2263.891	-1.48%	2302.966	2269.534	-1.47%
Apr.	2465.823	2432.391	-1.37%	2471.466	2438.034	-1.37%
May	2999.394	2965.962	-1.13%	3005.037	2971.605	-1.13%
Jun.	3657.180	3623.748	-0.92%	3662.823	3629.391	-0.92%
Jul.	4851.752	4818.320	-0.69%	4857.395	4823.963	-0.69%
Aug.	5319.537	5286.105	-0.63%	5325.18	5291.748	-0.63%
Sept.	5055.394	5021.962	-0.67%	5061.037	5027.605	-0.66%
Oct.	4738.966	4705.534	-0.71%	4744.609	4711.177	-0.71%
Nov.	3640.823	3607.391	-0.93%	3646.466	3613.034	-0.93%
Dec.	4905.323	4871.891	-0.69%	4910.966	4877.534	-0.69%

Table C1.2 Sensitivity of Decomposition Salmonellosis

Month	Forecast 2008			Forecast 2009		
	Data 93-06	Data 93-07	Senst (%)	Data 93-06	Data 93-07	Senst. (%)
Jan.	1636.770	1524.034	-7.40%	1594.011	1509.287	-5.61%
Feb.	1709.892	1565.510	-9.22%	1667.134	1550.764	-7.50%
Mar.	2084.422	1959.338	-6.38%	2041.664	1944.591	-4.99%
Apr.	2258.824	2097.523	-7.69%	2216.065	2082.776	-6.40%
May	2759.078	2608.082	-5.79%	2716.319	2593.336	-4.74%
Jun.	3448.781	3353.714	-2.83%	3406.022	3338.967	-2.01%
Jul.	4580.798	4486.279	-2.11%	4538.039	4471.533	-1.49%
Aug.	5110.750	4950.449	-3.24%	5067.992	4935.703	-2.68%
Sept.	4721.101	4683.018	-0.81%	4678.342	4668.272	-0.22%
Oct.	4545.185	4362.736	-4.18%	4502.426	4347.99	-3.55%
Nov.	3421.920	3262.540	-4.89%	3379.161	3247.793	-4.04%
Dec.	4586.690	4521.933	-1.43%	4543.932	4507.187	-0.82%

Table C1.3 Sensitivity of Holt-Winter's Salmonellosis

Month	**Forecast 2008**			**Forecast 2009**		
	Data 93-06	**Data 93-07**	**Senst (%)**	**Data 93-06**	**Data 93-07**	**Senst. (%)**
Jan.	2046.950	2187.529	6.43%	2079.757	2234.221	6.91%
Feb.	2124.456	2233.714	4.89%	2157.263	2280.406	5.40%
Mar.	2509.832	2636.357	4.80%	2542.639	2683.049	5.23%
Apr.	2681.149	2773.693	3.34%	2713.956	2820.385	3.77%
May	3193.841	3293.796	3.03%	3226.648	3340.488	3.41%
Jun.	3889.375	4041.991	3.78%	3922.182	4088.683	4.07%
Jul.	5063.161	5149.746	1.68%	5095.968	5196.438	1.93%
Aug.	5558.120	5642.917	1.50%	5590.927	5689.609	1.73%
Sept.	5189.427	5396.485	3.84%	5222.234	5443.176	4.06%
Oct.	5000.762	5069.071	1.35%	5033.57	5115.762	1.61%
Nov.	3885.963	3970.783	2.14%	3918.77	4017.475	2.46%
Dec.	5069.046	5243.875	3.33%	5101.853	5290.566	3.57%

Table C1.4 Sensitivity of ARIMA Salmonellosis

Month	**Forecast 2008**			**Forecast 2009**		
	Data 93-06	**Data 93-07**	**Senst (%)**	**Data 93-06**	**Data 93-07**	**Senst. (%)**
Jan.	1678.544	1693.071	0.86%	1678.558	1693.104	0.86%
Feb.	1965.624	1947.012	-0.96%	1965.666	1947.100	-0.95%
Mar.	1969.383	2016.797	2.35%	1969.399	2016.869	2.35%
Apr.	2692.268	2714.752	0.83%	2692.24	2714.683	0.83%
May	2657.578	2641.360	-0.61%	2657.604	2641.425	-0.61%
Jun.	3364.677	3470.293	3.04%	3364.646	3470.174	3.04%
Jul.	5020.563	5016.629	-0.08%	5020.595	5016.698	-0.08%
Aug.	4675.721	4679.952	0.09%	4675.72	4679.948	0.09%
Sept.	5655.426	5634.219	-0.38%	5655.587	5634.557	-0.37%
Oct.	5691.992	5681.368	-0.19%	5691.944	5681.259	-0.19%
Nov.	3489.962	3493.010	0.09%	3489.946	3492.974	0.09%
Dec.	5639.223	5624.434	-0.26%	5639.14	5624.257	-0.26%

Table C1.5 Sensitivity of Neural Network Salmonellosis

Month	Forecast 2008			Forecast 2009		
	Data 93-06	Data 93-07	Senst (%)	Data 93-06	Data 93-07	Senst. (%)
Jan.	1487.554	1883.876	21.04%	1401.877	1666.631	15.89%
Feb.	2333.572	2176.426	-7.22%	2405.847	2130.395	-12.93%
Mar.	2746.809	2736.039	-0.39%	2678.292	2581.257	-3.76%
Apr.	2123.442	2476.543	14.26%	2114.184	2488.578	15.04%
May	3233.492	3128.813	-3.35%	3389.266	3615.711	6.26%
Jun.	3583.734	4389.561	18.36%	3436.217	3958.325	13.19%
Jul.	4585.220	4541.067	-0.97%	4856.183	5171.531	6.10%
Aug.	6150.246	5574.243	-10.33%	6201.189	5298.844	-17.03%
Sept.	5337.590	4701.192	-13.54%	5051.095	4548.593	-11.05%
Oct.	3921.216	4814.853	18.56%	4106.460	4701.136	12.65%
Nov.	3937.339	3605.568	-9.20%	3912.961	3238.846	-20.81%
Dec.	5186.572	3742.482	-38.59%	5064.932	3612.997	-40.19%

C2. What If (Sensitivity) Analysis Result of Tuberculosis

Table C2.1 Sensitivity of Regression Tuberculosis

Month	Forecast 2008			Forecast 2009		
	Data 93-06	Data 93-07	Senst (%)	Data 93-06	Data 93-07	Senst. (%)
January	-23.989	48.143	149.83%	-100.651	-24.231	-315.38%
February	405.904	460.310	11.82%	329.242	387.936	15.13%
March	752.226	796.876	5.60%	675.563	724.502	6.75%
April	820.511	859.943	4.59%	743.849	787.569	5.55%
May	845.583	894.943	5.52%	768.920	822.569	6.52%
June	973.511	1015.543	4.14%	896.849	943.169	4.91%
July	855.797	901.743	5.10%	779.134	829.369	6.06%
August	894.297	940.076	4.87%	817.634	867.702	5.77%
September	791.726	845.010	6.31%	715.063	772.636	7.45%
October	854.154	900.476	5.14%	777.492	828.102	6.11%
November	723.654	779.676	7.19%	646.992	707.302	8.53%
December	2427.083	2426.676	-0.02%	2350.420	2354.302	0.16%

Table C2.2 Sensitivity of Decomposition Tuberculosis

Month	Forecast 2008			Forecast 2009		
	Data 93-06	**Data 93-07**	**Senst (%)**	**Data 93-06**	**Data 93-07**	**Senst. (%)**
January	390.954	227.432	-71.90%	360.742	173.143	-108.35%
February	643.112	361.183	-78.06%	593.093	273.218	-117.08%
March	841.896	464.748	-81.15%	775.989	349.215	-122.21%
April	869.101	471.281	-84.41%	800.618	351.646	-127.68%
May	891.688	480.920	-85.41%	820.960	356.199	-130.48%
June	961.845	508.333	-89.22%	885.045	373.591	-136.90%
July	892.212	465.145	-91.81%	820.495	339.066	-141.99%
August	908.447	469.378	-93.54%	834.932	339.211	-146.14%
September	852.770	429.392	-98.60%	783.292	307.497	-154.73%
October	886.862	439.246	-101.91%	814.113	311.532	-161.33%
November	807.687	395.004	-104.48%	740.976	277.302	-167.21%
December	1796.367	848.538	-111.70%	1646.969	589.255	-179.50%

Table C2.3 Sensitivity of Holt-Winter's Tuberculosis

Month	Forecast 2008			Forecast 2009		
	Data 93-06	**Data 93-07**	**Senst (%)**	**Data 93-06**	**Data 93-07**	**Senst. (%)**
January	445.865	467.937	4.72%	419.741	443.253	5.30%
February	725.867	751.192	3.37%	683.129	711.392	3.97%
March	955.700	986.230	3.10%	899.151	933.745	3.70%
April	985.133	1013.225	2.77%	926.554	959.063	3.39%
May	1019.213	1059.542	3.81%	958.306	1002.651	4.42%
June	1094.277	1132.438	3.37%	1028.556	1071.359	4.00%
July	1013.659	1050.888	3.54%	952.474	993.951	4.17%
August	1035.340	1081.082	4.23%	972.531	1022.244	4.86%
September	977.138	1012.997	3.54%	917.558	957.613	4.18%
October	1013.149	1051.164	3.62%	951.058	993.430	4.27%
November	923.821	965.082	4.28%	866.914	911.833	4.93%
December	2041.485	2107.085	3.11%	1915.081	1990.288	3.78%

Table C2.4 Sensitivity of ARIMA Tuberculosis

Month	**Forecast 2008**			**Forecast 2009**		
	Data 93-06	**Data 93-07**	**Senst (%)**	**Data 93-06**	**Data 93-07**	**Senst. (%)**
January	615.625	652.190	5.61%	593.536	635.403	6.59%
February	684.849	732.347	6.49%	663.319	712.687	6.93%
March	1332.413	1362.187	2.19%	1308.302	1342.137	2.52%
April	1453.616	1398.017	-3.98%	1435.339	1384.351	-3.68%
May	907.725	1007.076	9.87%	884.979	988.232	10.45%
June	1934.386	1900.321	-1.79%	1913.528	1883.961	-1.57%
July	1205.717	1168.970	-3.14%	1185.293	1150.378	-3.04%
August	1633.178	1720.047	5.05%	1610.218	1701.880	5.39%
September	881.893	1025.083	13.97%	861.532	1007.127	14.46%
October	1303.214	1386.662	6.02%	1279.862	1367.201	6.39%
November	1362.922	1304.543	-4.48%	1344.269	1290.923	-4.13%
December	2554.489	2709.159	5.71%	2534.060	2691.676	5.86%

Table C2.5 Sensitivity of Neural Network Tuberculosis

Month	**Forecast 2008**			**Forecast 2009**		
	Data 93-06	**Data 93-07**	**Senst (%)**	**Data 93-06**	**Data 93-07**	**Senst. (%)**
January	733.548	642.932	-14.09%	760.795	696.284	-9.27%
February	1091.572	817.229	-33.57%	1103.347	827.920	-33.27%
March	1000.541	1007.202	0.66%	1005.294	1016.179	1.07%
April	1078.629	980.305	-10.03%	1062.692	984.680	-7.92%
May	1058.650	1159.220	8.68%	1039.698	1165.370	10.78%
June	1133.220	1069.492	-5.96%	1113.412	998.991	-11.45%
July	1170.706	1014.166	-15.44%	1147.696	957.272	-19.89%
August	1190.171	1107.373	-7.48%	1176.088	1088.068	-8.09%
September	1079.647	1067.628	-1.13%	1066.586	1011.614	-5.43%
October	943.550	1116.316	15.48%	930.325	1125.435	17.34%
November	1078.699	1041.341	-3.59%	1069.786	993.849	-7.64%
December	1630.433	1711.533	4.74%	1533.755	1557.132	1.50%

APPENDIX D

SENSITIVITY PLOTS

D1. Sensitivity Analysis in 2008 and 2009 for Salmonellosis

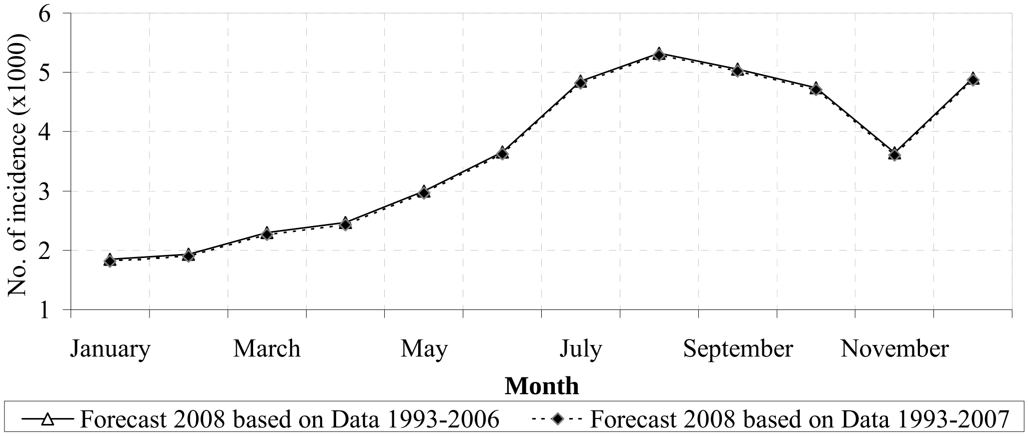


Figure D1.1 Sensitivity Analysis of Regression Forecast 2008 for Salmonellosis

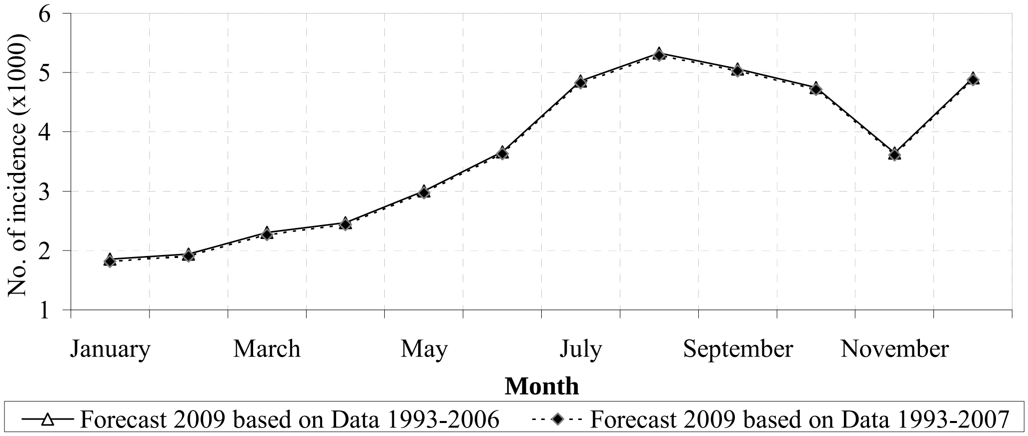


Figure D1.2 Sensitivity Analysis of Regression Forecast 2009 for Salmonellosis

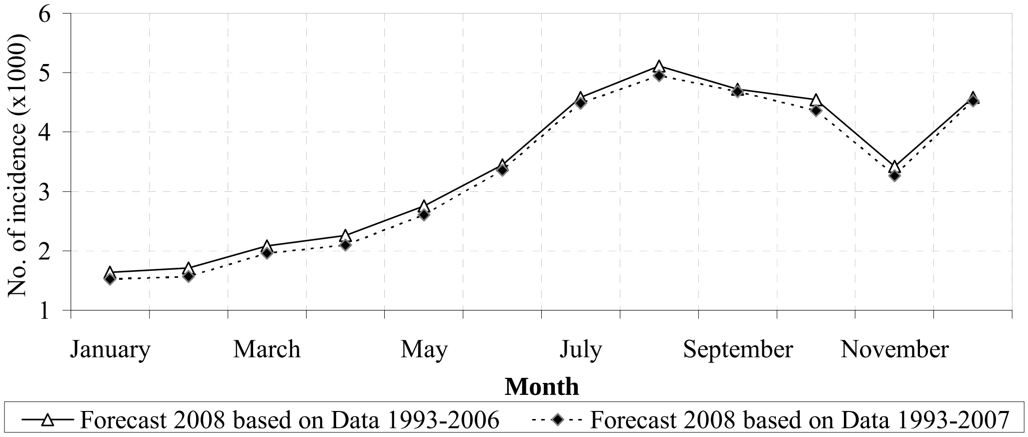


Figure D1.3 Sensitivity Analysis of Decomposition Forecast 2008 for Salmonellosis

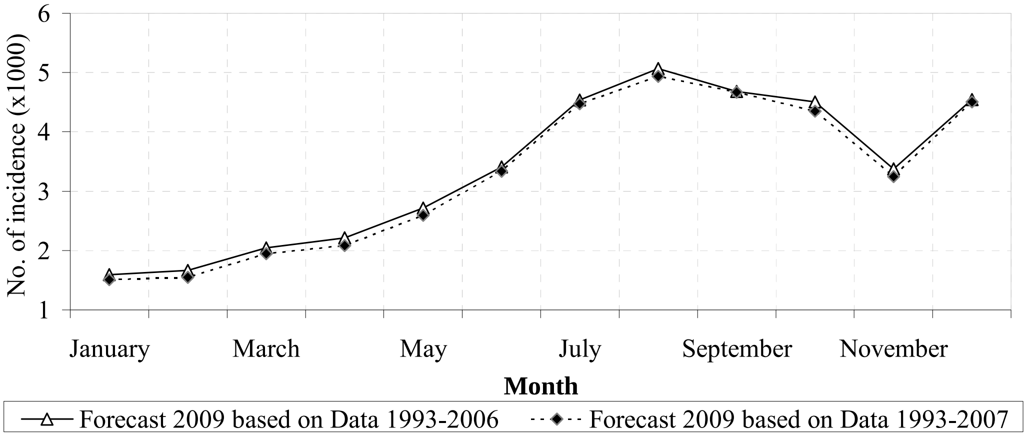


Figure D1.4 Sensitivity Analysis of Decomposition Forecast 2009 for Salmonellosis

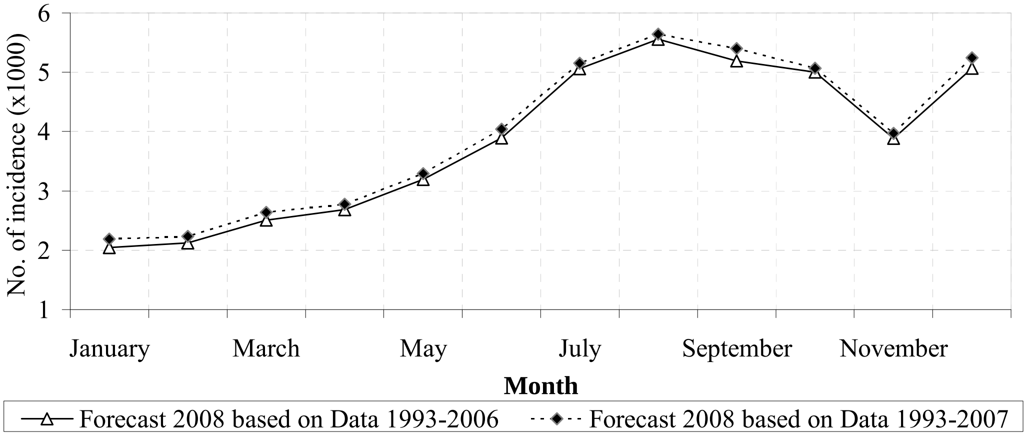


Figure D1.5 Sensitivity Analysis of Holt-Winter's Forecast 2008 for Salmonellosis

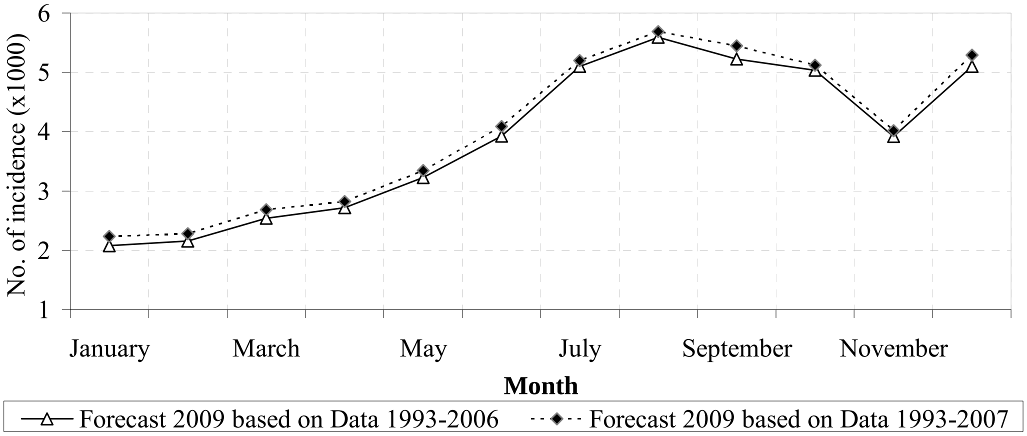


Figure D1.6 Sensitivity Analysis of Holt-Winter's Forecast 2009 for Salmonellosis

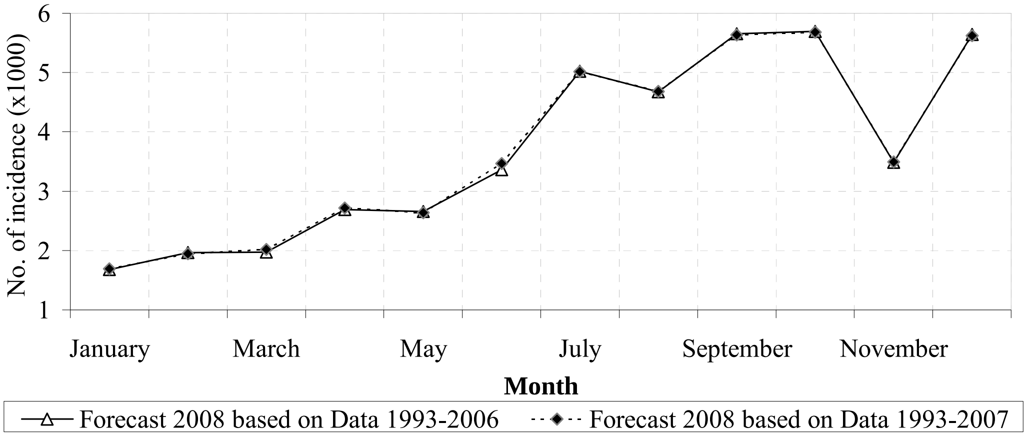


Figure D1.7 Sensitivity Analysis of ARIMA Forecast 2008 for Salmonellosis

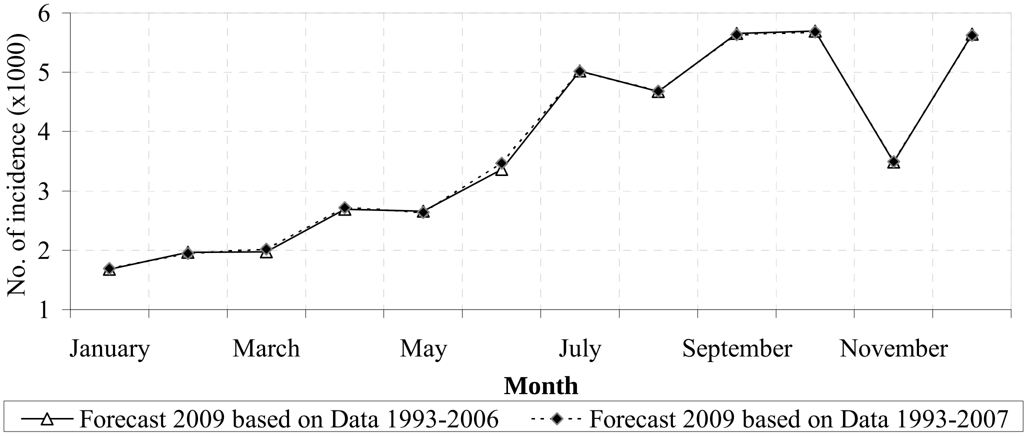


Figure D1.8 Sensitivity Analysis of ARIMA Forecast 2009 for Salmonellosis

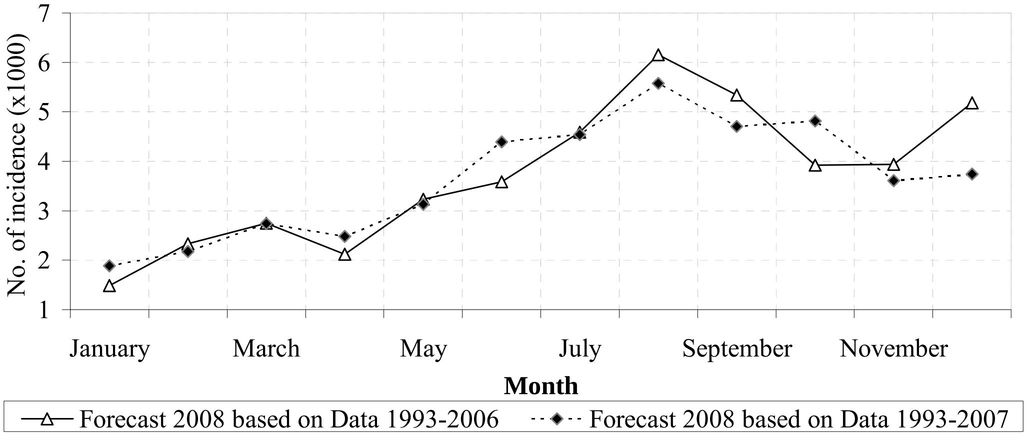


Figure D1.9 Sensitivity Analysis of Neural Network Forecast 2008 for Salmonellosis

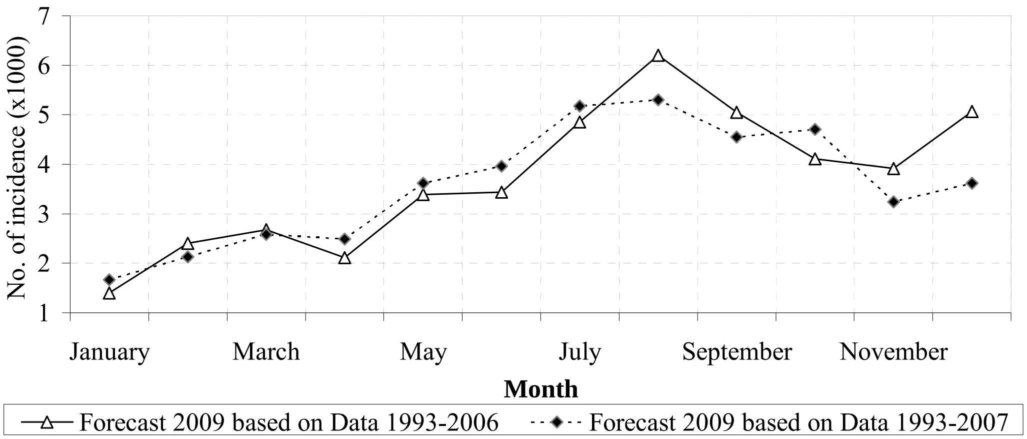


Figure D1.10 Sensitivity Analysis of Neural Network Forecast 2009 for Salmonellosis

C2. Sensitivity Analysis in 2008 and 2009 for Tuberculosis

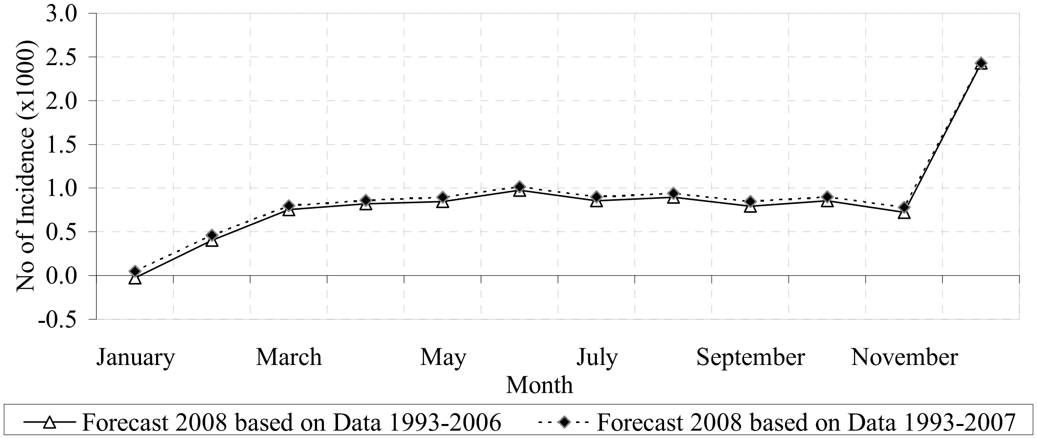


Figure D2.1 Sensitivity Analysis of Regression Forecast 2008 for Tuberculosis

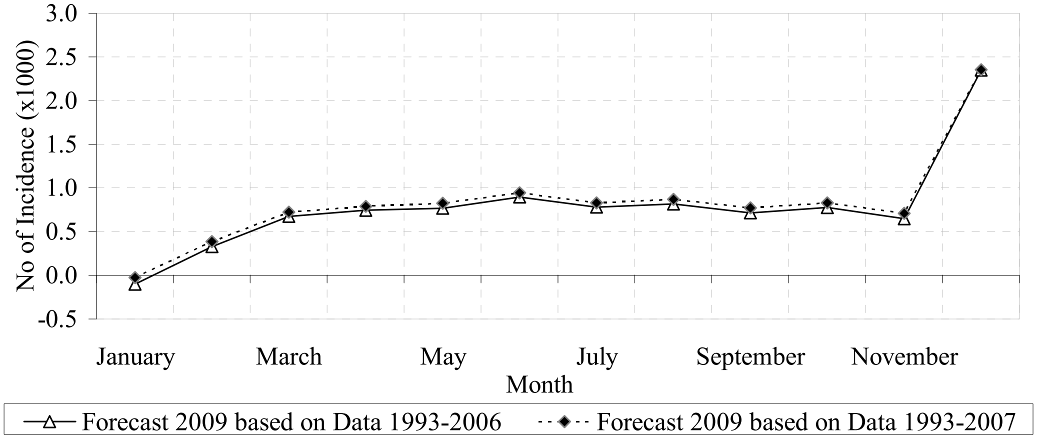


Figure D2. 2 Sensitivity Analysis of Regression Forecast 2009 for Tuberculosis

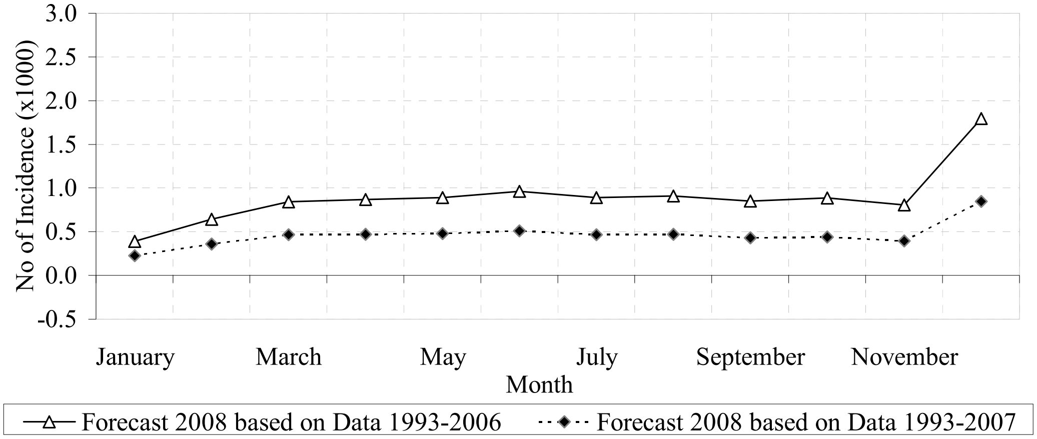


Figure D2.3 Sensitivity Analysis of Decomposition Forecast 2008 for Tuberculosis

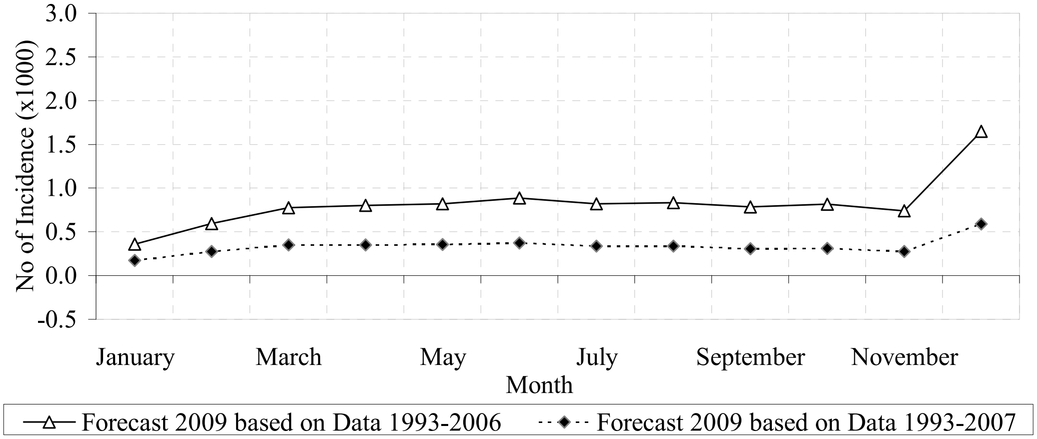


Figure D2.4 Sensitivity Analysis of Decomposition Forecast 2009 for Tuberculosis

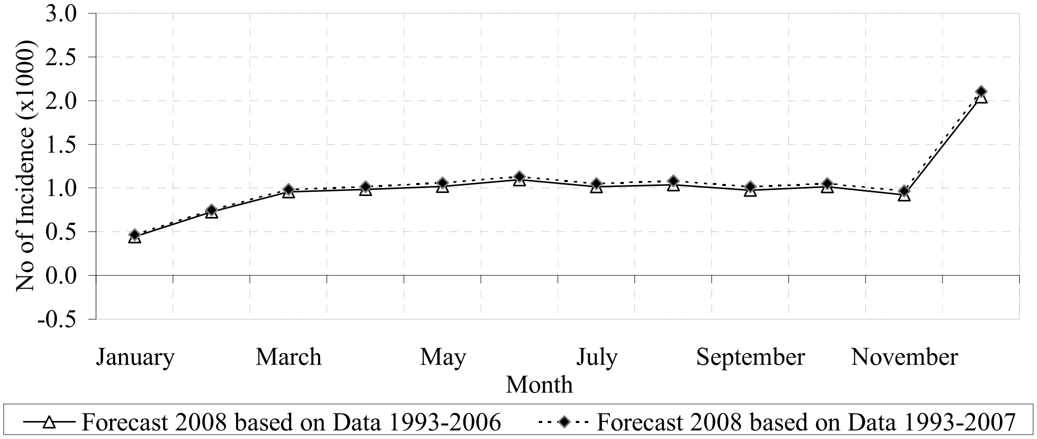


Figure D2.5 Sensitivity Analysis of Holt-Winter’s Forecast 2008 for Tuberculosis

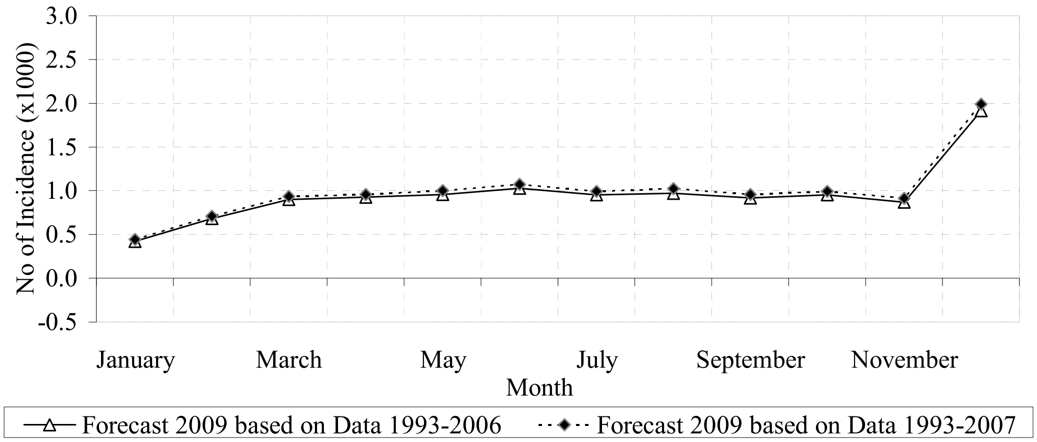


Figure D2.6 Sensitivity Analysis of Holt-Winter's Forecast 2009 for Tuberculosis

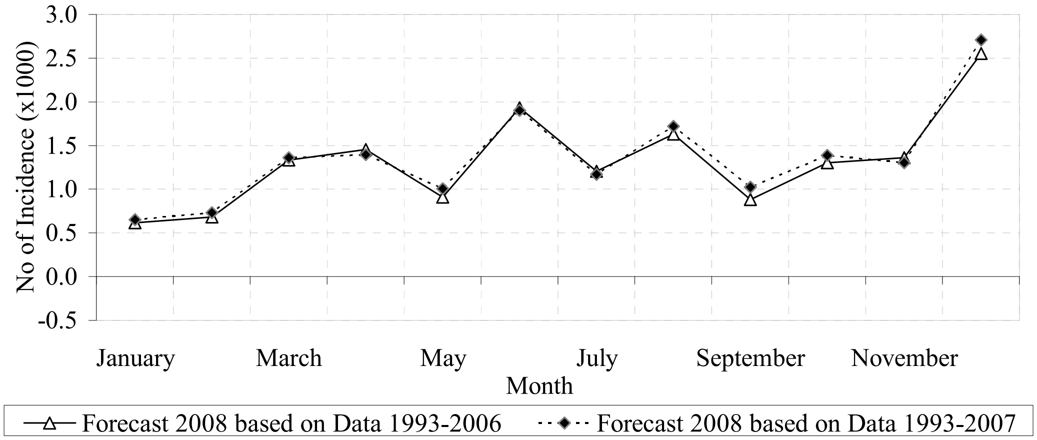


Figure D2.7 Sensitivity Analysis of ARIMA Forecast 2008 for Tuberculosis

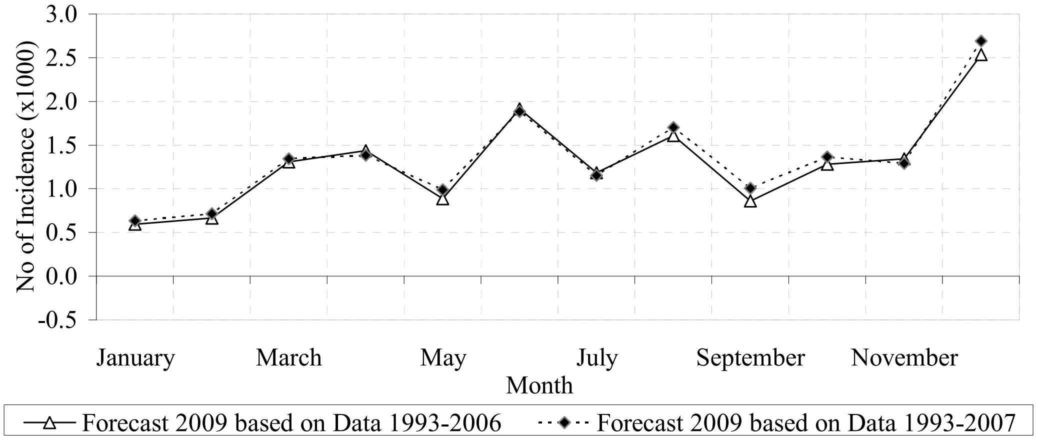


Figure D2.8 Sensitivity Analysis of ARIMA Forecast 2009 for Tuberculosis

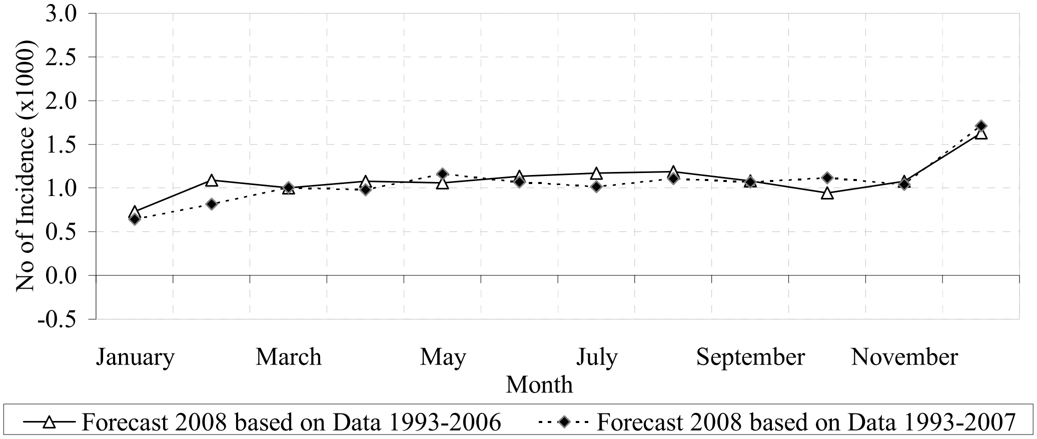


Figure D2.9 Sensitivity Analysis Neural Network Forecast 2008 for Tuberculosis

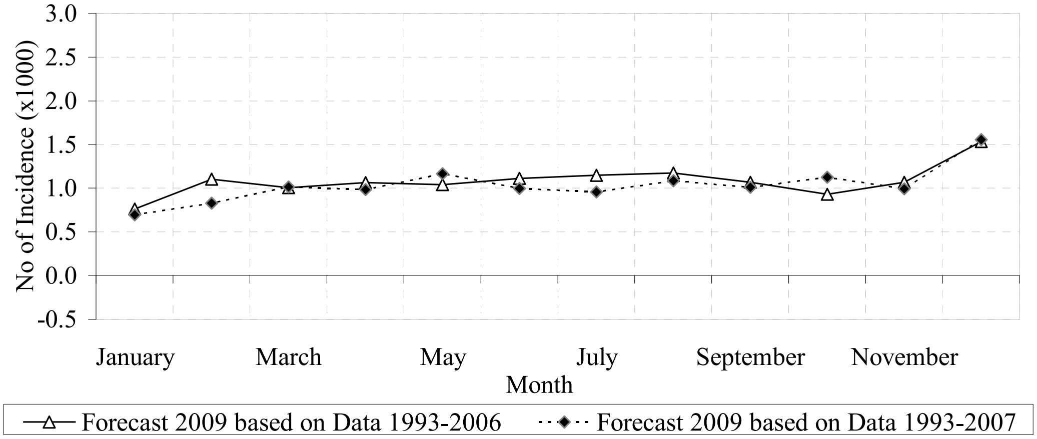


Figure D2.10 Sensitivity Analysis Neural Network Forecast 2009 for Tuberculosis