World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:1, No:8, 2007

Predicting DHF Incidence in Northern Thailand using Time Series Analysis Technique

S. Wongkoon, M. Pollar, M. Jaroensutasinee, and K. Jaroensutasinee

***Abstract*—**This study aimed at developing a forecasting model on the number of Dengue Haemorrhagic Fever (DHF) incidence in Northern Thailand using time series analysis. We developed Seasonal Autoregressive Integrated Moving Average (SARIMA) models on the data collected between 2003-2006 and then validated the models using the data collected between January-September 2007. The results showed that the regressive forecast curves were consistent with the pattern of actual values. The most suitable model was the SARIMA$(2,0,1)(0,2,0)_{12}$ model with a Akaike Information Criterion (AIC) of 12.2931 and a Mean Absolute Percent Error (MAPE) of 8.91713. The SARIMA$(2,0,1)(0,2,0)_{12}$ model fitting was adequate for the data with the *Portmanteau* statistic $Q_{20} = 8.98644$ ($\chi^2_{0.95} = 27.5871$, $P>0.05$). This indicated that there was no significant autocorrelation between residuals at different lag times in the SARIMA$(2,0,1)(0,2,0)_{12}$ model.

***Keywords*—**Dengue, SARIMA, Time Series analysis, Northern Thailand.

I. Introduction

DENGUE Fever (DF) and Dengue Haemorrhagic Fever (DHF) are caused by four antigenically distinct but related dengue virus serotypes transmitted primarily by *Aedes aegypti* and *Ae. albopictus*. DHF, the severe form of the disease, is endemic and frequently intensifies into epidemics in Southeast Asia, resulting in frequent hospitalisations and deaths [1], [2]. Recently, dengue has emerged as a substantial global health problem with increased incidences in new countries and tropical areas [3], [4].

DHF has been reported in Thailand since the late 1950s [5]-[7]. There has been an upward trend in the incidence of DHF, an acute and severe form of dengue virus infection. Since the first DHF epidemic outbreak in 1958 [8], epidemics have been reported from almost all most regions of the country. The Bureau of Epidemiology has reported that there have been several outbreaks reporting regularly in Thailand. The highest number of cases was reported in 1987 when the incidence rate was as high as 325 cases per 100,000 population based on the number of cases reported. The latest epidemic was in 1998 when the incidence rate was as high as 211 cases per 100,000 populations. This was the second highest incidence rate in the 40 year history of DHF outbreaks [9].

The number of DHF incidence in Thailand from January-September 2007 was 47,454 cases (75.53 cases per 100,000 populations, Fatality rate was 0.13). In Northern Thailand, the number of DHF incidence was 6,713 cases (56.46 cases per 100,000 populations, Fatality rate was 0.10) [10]. In June 2007, the number of DHF incidence in Chiang Rai province, Northern Thailand was highest in Thailand (464 cases) [10].

Time series analysis has been used extensively in the assessment of the health sciences [11]-[13]. In health science research, there is often an obvious time lag between response and explanatory variables [14]. Some studies approach this by examining models with simultaneous multiple lags of the explanatory variables [15]. However, serial correlation between these variables may produce unstable estimates [14].

Forecasting DHF incidence in Northern Thailand by using time series models would provide useful information. The main characteristic of the time series modelling is that time series analysis only models the relationship between the observed DHF incidence at time t (y_t) from the past observations ($y_1, y_2, \ldots, y_{t-1}$), without using any other variables [16]. This study aimed at developing time series models to forecast the monthly DHF incidence in Northern Thailand, based on reported incidence available from 2003-2006 and then validated the models using the data collected between January-September 2007. This forecasting offers the potential for improved contingency planning of public health intervention in Northern Thailand.

II. Materials and Methods

A. Study Site

Northern Thailand covers an area of 170,000 km^2 and is located at latitude 16-21 °N and longitude 97-101 °E. This area is mostly high mountainous and covered with forest with several flat river basins, bordering on the territories of Laos and Myanmar. The region divided into 17 provinces with a local population of 11,842,299 and a density of 69.7 people/km^2 [17]. The climate of Northern Thailand is dominated by two tropical monsoons: southwest and northeast monsoon. Southwest monsoon starts in May whereas northeast monsoon begins in November. As a result of these two monsoons, the seasonal weather for Northern Thailand consists of three seasons: summer season (February-May),

Manuscript received October 15, 2007. This work was supported in part by the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/0201/2548), CXKURUE, the Institute of Research and Development, Walailak University, The GLOBE program, IPST.

S. Wongkoon is with School of Science, Walailak University, 222 Thaiburi, Thasala District, Nakhon Si Thammarat 80161, Thailand (phone: +66 75 672 048; Fax: +66 75 672 038; e-mail: swongkoon@gmail.com).

M. Pollar is with School of Science, Walailak University, 222 Thaiburi, Thasala District, Nakhon Si Thammarat 80161, Thailand (phone: +66 75 672 075; Fax: +66 75 672 038; e-mail: m_pollar@hotmail.com).

M. Jaroensutasinee is with School of Science, Walailak University, 222 Thaiburi, Thasala District, Nakhon Si Thammarat 80161, Thailand (phone: +66 75 672 005; Fax: +66 75 672 004; e-mail: jmullica@gmail.com).

K. Jaroensutasinee is with School of Science, Walailak University, 222 Thaiburi, Thasala District, Nakhon Si Thammarat 80161, Thailand (phone: +66 75 672 005; Fax: +66 75 672 004; e-mail: krisanadej@gmail.com).

rainy season (May-October) and winter season (October-February) [18].

B. Data Collection

The monthly DHF incidence in Northern Thailand from January 2003-September 2007 was provided by the Department of Medical Science, Ministry of Public Health.

C. Statistical Analysis

For modelling, the monthly DHF incidence in Northern Thailand was divided into two parts: (1) DHF data observed during January 2003-December 2006 were used for developing the models, and (2) DHF data during January-September 2007 were used for validating the time series models. The original time series of monthly DHF incidence in Northern Thailand at time t (y_t) was (y_1, y_2, …, y_{t-1}) (Fig. 2). The monthly DHF incidences were transformed to become stationary input series with respect to yearly periodicity by seasonally differencing before being modelled (Fig. 3).

Seven main parameters were selected when fitting the SARIMA $(p,d,q)(P,D,Q)_s$ model: the order of autoregression (p) and seasonal autoregression (P), the order of integration (d), seasonal intergration (D), the order of moving average (q), seasonal moving average (Q), and the length of seasonal period (s). All analyses were performed using the *Mathematica* software with Time Series application.

There were five steps to perform in order to obtain the coefficients in Table I [19]:

(a) *Series Stationarity*. ARMA modelling was used correctly if the time series was Wide Sense Stationary (WSS). Exponential decreasing of autocorrelation function across lags was a strong indication of time series stationarity [19]. Therefore, the first step was to transform the time series data by seasonal differencing in order to achieve time series stationarity.

(b) *Identification*. We estimated the order of AR and MA using autocorrelation (ACF) and partial autocorrelation functions (PACF).

(c) *Optimisation*. The model coefficients were calculated using a training group by means of the autocorrelation computed in the previous step.

(d) *Selection*. The most suitable models were chosen based on their adequate predictions. In order to evaluate models, data were split into two groups: training and validation. The training group was used to build the time series model, and the validation group was used to evaluate the time series model. Akaike Information Criterion (AIC) based on information theory was used to achieve a trade-off between an adequate prediction and a few number of parameters [20].

(e) *Residues*. Autocorrelation and the mean absolute percent error (MAPE) of the residues were calculated to test whether they were statistically relevant (1).

$$\text{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right| \quad (1)$$

Fig. 1 Map of Northern Thailand

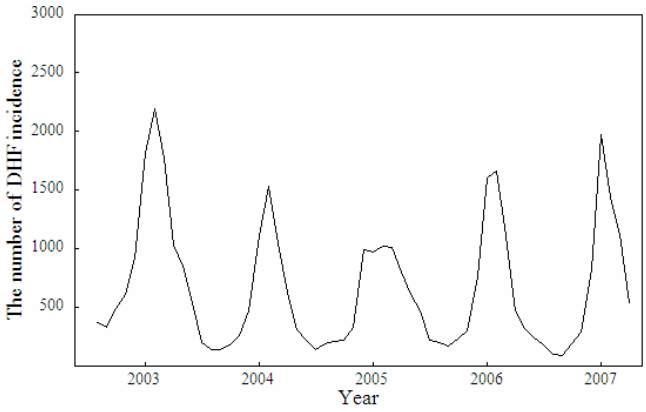

Fig. 2 The number of DHF incidences in Northern Thailand from January 2003-September 2007

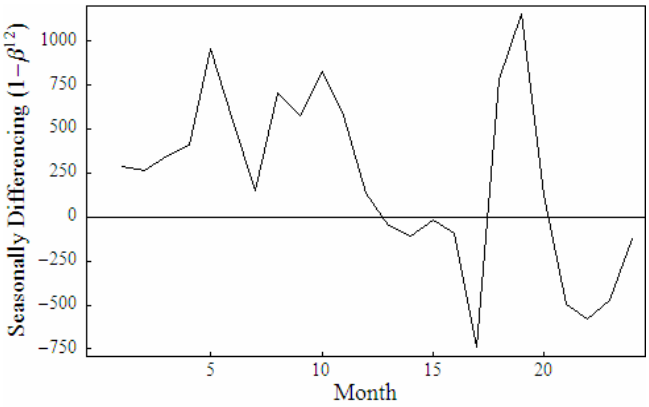

Fig. 3 Seasonally differencing 2nd order with 12 months of DHF incidence in Northern Thailand from January 2003-December 2006

III. Results and Discussion

ACF and PACF of DHF time series were exponentially tailing off (Fig. 4, 5). These results indicate the 2nd order of differenced time series model.

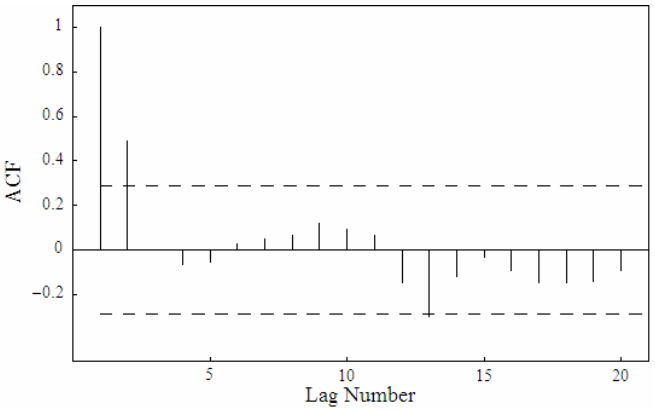

Fig. 4 ACF of DHF incidence in Northern Thailand between January 2003-December 2006 (--- represented 95% upper and lower confidence intervals)

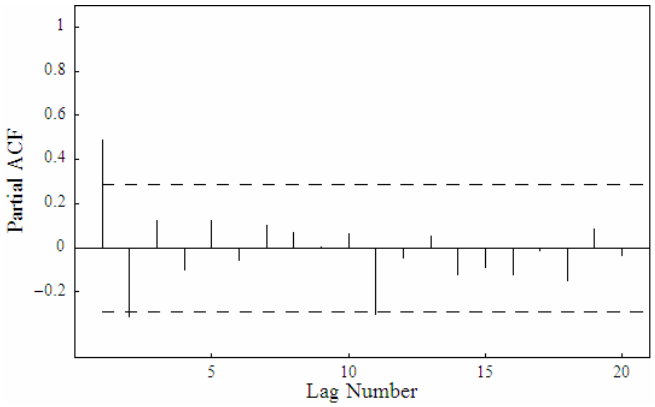

Fig. 5 PACF of DHF incidence in Northern Thailand between January 2003-December 2006 (--- represented 95% upper and lower confidence intervals)

There were six suitable time series models for forecasting the number of DHF incidence in Northern Thailand (Table I). We selected time series models based on the lowest AIC and MAPE values (Table I). SMA(1) model had the lowest AIC and SARIMA(2,0,1)(0,2,0)$_{12}$ model had lowest MAPE. The ACF of residuals at different lag times in SMA(1) model was differed from zero (Fig. 6a). ACF of residuals at different lag times in SARIMA(2,0,1)(0,2,0)$_{12}$ model was not differed from zero (*Portmanteau* statistic Q_{20} = 8.98644, $\chi^2_{0.05}$ = 27.5871, P>0.05). The graphic analysis of residuals showed that the residuals in the model appeared to fluctuate randomly around zero with no obvious trend in variation as the predicted incidence values increased (Fig. 6b). This indicates that the most suitable model for predicting DHF incidence in Northern Thailand was the SARIMA(2,0,1)(0,2,0)$_{12}$ model:

$$(1-B)(1-B^{12})Y_t = 0.18 + (1-0.06B)(1-0.61B^{12})Z_t \quad (2)$$

The observed and predicted DHF cases from January-September 2007 matched reasonably well. The predicted DHF cases for the year 2007 increased and reached a maximum predicted case in July 2007 (Fig. 7).

TABLE I
MAPE AND AIC OF TIME SERIES MODELS

Models	MAPE	AIC
SMA(1)	9.68048	12.2222
SARIMA(1,0,1)(0,2,0)$_{12}$	9.51984	12.2546
SMA(2)	9.58155	12.2937
SARIMA(2,0,1)(0,2,0)$_{12}$	8.91713	12.2931
SARIMA(1,0,2)(0,2,0)$_{12}$	9.45025	12.3351
SARIMA(2,0,2)(0,2,0)$_{12}$	9.50031	12.3746

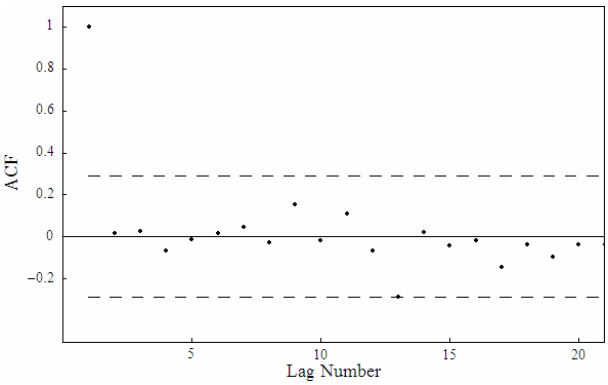

Fig. 6 (a) The correlation function of residuals from SMA(1) model (--- represented 95% upper and lower confidence intervals)

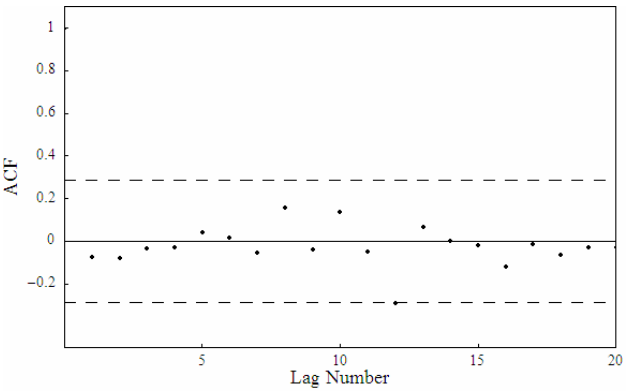

Fig. 6 (b) The correlation function of residuals from SARIMA(2,0,1)(0,2,0)$_{12}$ model (--- represented 95% upper and lower confidence intervals)

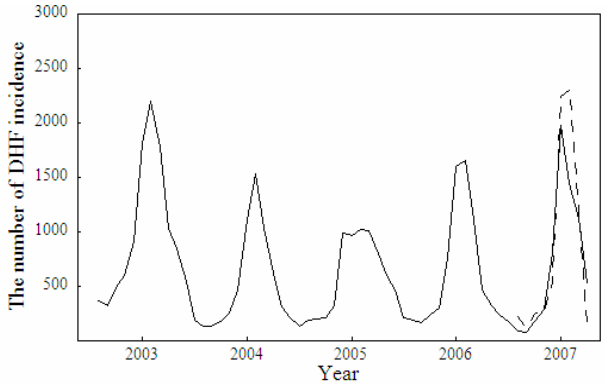

Fig. 7 The number of DHF incidences in Northern Thailand from January 2003-September 2007. ─ represented actual data, --- represented predicted data

The results of this study indicate that SARIMA model allows for more complex description of the seasonality and autocorrelation structure of the series and appears to be suitable in predicting the number of DHF incidences in Northern Thailand. The SARIMA model fits the observed DHF incidences well. The residuals did not deviate significantly from a zero mean white noise process.

There has been a remarkable advance in modelling approaches on public health and epidemiology [21]-[23]. The development of mathematical models has been very useful in the control and prevention of infectious diseases. Some models have been developed to predict the likelihood of vector-disease epidemic using weather and environmental data [24]-[26]. It is crucial to use adequate research methodology in the assessment of possible impacts of environmental variability on disease transmission. Recently, increasing attention has focused on the use of the Box-Jenkins modelling strategy to construct (SARIMA) models for vector-borne disease [27]-[29].

The SARIMA model approach has several advantages over others such as moving average, exponential smoothing, neural network and fuzzy logic, in particular, its forecasting capability and its richer information on time-related changes [30], [31]. The steps of model identification, parameter estimation, and diagnostic checking are performed as recommended [30], [32].

SARIMA modelling is useful for interpreting and applying surveillance data in disease control and prevention [27]-[29], [33]. However, one of the most important for the SARIMA modelling approach is the necessity of large amount of data (i.e., a minimum of 50 observations) to build reasonable SARIMA model [34]. In this study we used 48 observations to develop the model. If more data are available, this SARIMA model can be improved.

ACKNOWLEDGMENTS

This work was supported in part by the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/0201/2548), CXKURUE, the Institute of Research and Development, Walailak University, the GLOBE program, IPST. We thank Robert James Santos for his comments on the previous versions of this manuscript. We also thank Miss Ruthairat Sritommarat for providing the DHF incidence in Thailand from Department of Medical Science, Ministry of Public Health.

REFERENCES

[1] World Health Organization. Dengue haemorrhagic fever: diagnosis, treatment, prevention and control (2nd edition). Geneva: World Health Organization; 1997.

[2] S. Anuradha, N. P. Singh, S. N. Rizvi, S. K. Agarwal, R. Gur, and M. D. Mathur, "The 1996 outbreak of dengue hemorrhagic fever in Delhi, India," *Southeast Asian J Trop Med Public Health*, vol. 29, pp. 503-506, 1998.

[3] J. G. Rigau-Perez, D. J. Gubler, A. V. Vorndam, and G. G. Clark, "Dengue: literature review and case study of travelers from the United States 1986–1994," *J Travel Med*, vol. 4, pp. 65-71, 1997.

[4] D. J. Gubler, "The global pandemic of dengue/dengue haemorrhagic fever: current status and prospects for the future," *Ann Acad Med Singapore*, vol. 27, pp. 227-234, 1998.

[5] W. M. Hammon, A. Rudnick, and G. E. Sather, "Viruses associated with epidemic hemorrhagic fevers of the Philippines and Thailand," *Science*, vol. 131, pp. 1102-1103, 1960.

[6] S. B. Halstead, "Mosquito-borne haemorrhagic fevers of south and southeast Asia," *Bull World Health Organ*, vol. 35, pp. 3-15, 1966.

[7] K. Ungchusak, and P. Kunasol, "Dengue haemorrhagic fever in Thailand, 1987," *Southeast Asian J Trop Med Public Health*, vol. 19(3), pp. 487-490, 1988.

[8] P. Barbazan, S. Yoksan, and J. P. Gonzalez, "Dengue hemorrhagic fever epidemiology in Thailand: description and forecasting of epidemics," *Microbes Infect*, vol. 4(7), pp. 699-705, 2002.

[9] Bureau of Epidemiology, "DHF Situation in Thailand," Ministry of Public Health, Thailand, 2002.

[10] Department of Medical Science, "DHF Situation in Thailand," Ministry of Public Health, Thailand, 2007.

[11] C. Bowie, and D. Prothero, "Finding cases of seasonal diseases using time series analysis," *Int. J. Epidemiol*, vol. 10, pp. 87-92, 1981.

[12] R. Catalano, and S. Serxner, "Time series designs of potential interest to epidemiologists," *Am. J. Epidemiol*, vol. 26, pp. 724-731, 1987.

[13] U. Helfenstein, "The use of transfer function models, intervention analysis and related time series methods in epidemiology," *Int. J. Epidemiol*, vol. 20, pp. 808-815, 1991.

[14] J. Schwartz, C. Spix, G. Touloumi, L. Bacharova, T. Barumamdzadeh, A. Tertre, T. Pickarksi, A. Leon, A. Ponka, G. Rossi, M. Saez, and J. Schouten, "Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions," *J. Epidemiol. Community Health*, vol. 50(s), pp. s3-s11, 1996.

[15] J. Schwartz, "The distributed lag between air pollution and daily deaths," *Epidemiology*, vol. 11, pp. 320-326, 2000.

[16] T. Slini, K. Karatzas, and N. Moussiopoulos, "Statistical analysis of environmental data as the basis of forecasting: an air quality apllication," *Sci Total Environ,* vol. 288(3), pp. 227-237, 2002.

[17] National Statistical Office, "The population and area data in Thailand," Avalable online http://portal.nso.go.th, 2007

[18] Thai Meteorological Department, "The climatic data in Thailand," Avalable online : http://www.tmd.go.th, 2007.

[19] S. Makridakis, S. C. Wheelwrigth, and R. J. Hyndman, "Forecasting: methods and applications," *Prentice Hall*, 1998.

[20] A. S. Weigend, and N. A. Gershenfeld, "Time series prediction: forecasting the future and understanding the past," *Addison-Wesley*, 1996.

[21] A. Guisan, and N. Zimmermann, "Predictive habitat distribution models in ecology," *Ecol. Model*, vol. 135, pp. 147-186, 2000.

[22] S. Jørgensen, "25 years of ecological modelling by ecological modelling," *Ecol. Model*, vol. 126, pp. 95-99, 2000.

[23] S. Jørgensen, "Ecological modelling: editorial overview 2000-2005," *Ecol. Model*, vol. 188, pp. 137-144, 2005.

[24] R. Woodruff, C. Guest, M. Garner, N. Becker, J. Lindesay, T. Carvan, and K. Ebi, "Predicting Ross River virus epidemics from regional weather data," *Epidemiology*, vol. 13, pp. 384-393, 2002.

[25] H. Teklehaimanot, M. Lipsitch, A. Teklehaimanot, and J. Schwartz, "Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I patterns of lagged weather effects reflect biological mechanisms," *Malar. J*, vol. 3, pp. 41. 2004.

[26] H. 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," *Malar. J*, vol. 3, pp. 44. 2004.

[27] K. Linthicum, A. Anyamba, C. Tucker, P. Kelley, M. F. Myers, and C. Peters, "Climate and satellite indicators to forecast Rift Valley fever epidemics in Kenya," *Science*, vol. 285, pp. 347-348, 1999.

[28] T. Abeku, S. deVlas, G. Borsboom, A. Teklehaimanot, A. Kebede, D. Olana, G. van Oortamrssen, and J. Habbema, "Forecasting malaria incidence from historical morbidity patterns in epidemic-prone areas of Ethiopia: a simple seasonal adjustment method performs best," *Trop. Med. Int. Health*, vol. 7, pp. 851-857, 2002.

[29] W. Hu, N. Nicholls, M. Lindsay, P. Dale, A. McMichael, J. Mackenzie, and S. Tong, "Development of a predictive model for Ross River virus disease in Brisbane," *Am. J. Trop. Med. Hyg*, vol. 71, pp. 129-137, 2004.

[30] G. Box, and G. Jenkins, "Time-series Analysis: Forecasting and Control ," *Holden-Day (Maidenhead McGraw-Hill)*, 1970.

[31] K. Yurekli, A. Kurunc, and F. Ozturk, "Application of linear stochastic models to monthly flow data of Kelkit Stream," *Ecol. Model*, vol. 183, pp. 67–75, 2005.
[32] U. Helfenstein, "Box-Jenkins modelling of some viral infectious diseases," *Stat. Med*, vol. 5, pp. 37-47, 1986.
[33] R. Allard, "Use of time-series analysis in infectious disease surveillance," *Bull. World Health Organ*, vol. 76, pp. 327-333, 1998.
[34] W. Wei, "Time Series Analysis," *Addison-Wesley Publishing Company Inc.*, Now York, 1990.