Computational and Mathematical Methods in Medicine
IMPACT FACTOR 0.94

Comput Math Methods Med. 2012; 2012: 758674.
Published online 2012 Mar 8. doi: 10.1155/2012/758674

PMCID: PMC3310403

Comparing Statistical Models to Predict Dengue Fever Notifications

Arul Earnest,[1,2,*] Say Beng Tan,[1] Annelies Wilder-Smith,[3,4] and David Machin[5,6]

[1]Centre for Quantitative Medicine, Office of Clinical Sciences, Duke-NUS Graduate Medical School Singapore, Singapore 169857
[2]Tan Tock Seng Hospital, Singapore 308433
[3]Institute of Public Health, University of Heidelberg, Germany
[4]National University of Singapore, Singapore 119077
[5]University of Leicester, UK
[6]University of Sheffield, UK
*Arul Earnest: arul.earnest@duke-nus.edu.sg

Academic Editor: Chris Bauch

Received 2011 Sep 16; Revised 2011 Dec 7; Accepted 2011 Dec 8.

Abstract

Go to: Go to:

Dengue fever (DF) is a serious public health problem in many parts of the world, and, in the absence of a vaccine, disease surveillance and mosquito vector eradication are important in controlling the spread of the disease. DF is primarily transmitted by the female *Aedes aegypti* mosquito. We compared two statistical models that can be used in the surveillance and forecast of notifiable infectious diseases, namely, the Autoregressive Integrated Moving Average (ARIMA) model and the Knorr-Held two-component (K-H) model. The Mean Absolute Percentage Error (MAPE) was used to compare models. We developed the models using used data on DF notifications in Singapore from January 2001 till December 2006 and then validated the models with data from January 2007 till June 2008. The K-H model resulted in a slightly lower MAPE value of 17.21 as compared to the ARIMA model. We conclude that the models' performances are similar, but we found that the K-H model was relatively more difficult to fit in terms of the specification of the prior parameters and the relatively longer time taken to run the models.

1. Introduction

Go to: Go to:

The incidence of dengue fever (DF) has grown dramatically around the world in recent decades, with some 2.5 billion people now at risk of the disease [1]. Dengue haemorrhagic fever (DHF) is a potentially lethal complication, with an estimated 500 000 people requiring hospitalization each year, a very large proportion of whom are children. About 2.5% of those affected die [1].

DF is a viral vector-borne disease that is common in the tropics and subtropics and is primarily spread by the female *Aedes aegypti* mosquito. Mosquito vector control is important in restricting its spread. It has been found that controlling the vector population before disease is detected reducing transmission with a reduction of the *Aedes aegypti* population in a 3-month period, from 16% to 2%, as measured by the premises index [2]. However, predicting the incidence of vector-borne diseases like DF remains difficult, as DF shows strong variations over time [3–5]. In Singapore, seasonal trends are seen with peaks occurring generally in June or September. DF is characterized by both epidemic peaks that appear every 3–5 years, as well as seasonal oscillations within a year. Possible reasons for changes in outbreak patterns include change in number of infections due to interventions to eradicate the mosquitoes, as well as change in the number of people who are susceptible to the disease through prior infections [6]. Seasonal trends in DF can be caused by several factors, including climatic variables such as temperature and precipitation [7–10].

Autoregressive Integrated Moving Average (ARIMA) models have been used in applications such as the assessment of seasonal variation in selected medical conditions [11], and as a surveillance tool for outbreak detection [12]. ARIMA (AR, D, MA) models make use of previous observations to make predictions of future values using lag parameter values. Lags of the differenced series appearing in the forecasting equation are termed Auto Regressive (AR), those of the forecast errors, Moving Average (MA), and a time series that needs to be differenced to achieve stationarity, Differenced (D). The prediction process uses constantly updated information (in our example weekly DF cases) to predict the course of dengue in subsequent weeks.

Time series analysis of infectious diseases within the Bayesian framework has been considered in some studies [13–16]. One such example demonstrated that *Klebsiella pneumoniae* is related to the quantity of a third-generation antibiotic use (cephalosporin) in a hospital, with a lag of three months [17]. Others included a Knorr-Held (K-H) two-component model to incorporate both seasonal and epidemic characteristics of notifiable infectious diseases [15], as well as a Bayesian hierarchical time series model to detect outbreaks of Rubella and Salmonella infections [14].

Studies have compared ARIMA models with dynamic models for infectious diseases (fitted via maximum likelihood methods) [18, 19]. However, to the best of our knowledge, a direct comparison between the single-component (ARIMA) and two-component (K-H) models has not been undertaken.

2. Methods

Go to: Go to:

The purpose of this paper is to compare the two-component K-H with the single-component ARIMA model in predicting weekly DF notifications. Different formulations of models within each type are compared, together with a sensitivity analysis of the K-H model, fitted within a Bayesian framework.

2.1. Data

The Singapore Infectious Diseases Act (1977) requires medical practitioners to notify all cases of DF to the Ministry of Health (MoH) within 24 hours. We obtained data from the published “Weekly Infectious Disease Bulletin”, available from the MoH website which uses the World Health Organization 2009 criteria for DF which is also detailed there [20]. All notified and registered DF cases were laboratory confirmed, with laboratory assays from Polymerase Chain Reaction (PCR) and/or NS1 antigen (in the first 5 days of illness) and/or a positive Dengue Immunoglobulin M after day 5 of illness.

We studied weekly DF notifications in Singapore till June 2008. Data from January 2001 to December 2006 was used to estimate the model parameters. Thereafter, we performed external validation of the models using data from January 2007 to June 2008.

2.2. ARIMA Model

If f_T represents the number of cases of DF in week T, then AR relates this observation to an earlier f_{T-J}, where $J = 1,2,\ldots, T-1$. MA relates the error (defined as the difference between the observed, f, and that predicted, F, notifications) at week T to week $(T-K)$, where $K = 1,2,\ldots$. D allows the differenced series, $\Delta T = (f_T - f_{T-L})$, to be modelled in the event of nonstationarity in the time series, where $L = 0,1, 2,\ldots$. Here J, K, and L are the “orders” of the respective ARIMA components. Partial autocorrelation (PAC) and autocorrelation (AC) plots are used to determine J and K, respectively.

We describe the ARIMA (3,1,1) model equation used in our analysis. The number of cases of DF at week T is denoted as f_T, where T is the first week for which DF is to be predicted

$$F_T = \mu + f_{T-1} + \varphi_1 f_{T-1} + \varphi_2 f_{T-2} + \varphi_3 f_{T-3} + \theta \varepsilon_{T-1} + \varepsilon_T, \qquad (1)$$

where F_T is the predicted number of DF cases for week T, and f_{T-1}, f_{T-2}, and f_{T-3} are the DF counts in the three immediate preceding weeks, termed lag 1, lag 2, and lag 3, respectively, and ε_T, ε_{T-1} are the error term at time T and $T-1$, respectively. In essence, we used observed values up till time $T-1$ to predict for dengue fever cases at week T. μ is a constant and φ_1, φ_2, and φ_3 are the coefficients for the three autoregressive terms in the model, θ is the first order moving average parameter and these are estimated within Stata V11.0 [21] via full or unconditional maximum likelihood estimates. For the ARIMA models, we used the Mean Absolute Percentage Error (MAPE) described below to compare predictive accuracy of the models.

2.3. Two-Component K-H Model

The K-H model distinguishes between the endemic, x, and epidemic, y, components of DF such that the number of cases observed $f_T = x_T + y_T$ and the corresponding prediction model is formulated as

$$F_T = X_T + Y_T. \qquad (2)$$

Here X_T and Y_T have independent Poisson distributions with a composite parameters ($\omega_T v_T$) and ($\omega_T \lambda_T F_{T-1}$), in which ω_T handles over dispersion, hence F_T is also Poisson with parameter $\omega_T[v_T + \lambda_T F_{T-1}]$. This in turn corresponds to a negative binomial distribution with dispersion parameter ψ. The mixing parameter, ω_T, is assumed to have a Gamma distribution with parameters ($\psi + F_T$) and ($\psi + v_T + \lambda_T F_{T-1}$).

The endemic parameter, v_T, is modelled as a harmonic wave (to handle strong seasonality inherent in infectious disease surveillance data) with

$$\log \nu_T = \gamma_0 + \gamma_1 \sin\left(\frac{2\pi T}{52}\right) + \gamma_2 \cos\left(\frac{2\pi T}{52}\right), \qquad (3)$$

see [15], where $2\pi/52$ is the base frequency of the curve, which is suitable for weekly data and γ_0 is a constant. The logarithmic transformation is necessary to ensure stationarity in the variance of the series.

The epidemic component is derived from the parameter sequence $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_n)$, which is assumed to be a piecewise constant [15] with unknown number of location K and unknown location of the changepoints $\theta_1 < \cdots < \theta_K$, that is,

$$\lambda_T = \begin{cases} \lambda^{(1)}, & \text{if } T = 1,2,\ldots,\theta_1, \\ \lambda^{(k)}, & \text{if } T = \theta_{k-1}+1,\ldots,\theta_k, \\ \lambda^{(K+1)}, & \text{if } T = \theta_K + 1,\ldots,n, \end{cases} \qquad (4)$$

where $\theta_1 < \theta_2 < \cdots < \theta_K$ are the K unknown changepoints, such that $\theta_k \in \{1,2,\ldots, n-1)$ for all $k \in$

(1,2,…, K). For $K = 0$, there is no changepoint and $\lambda_T = \lambda^{(1)}$ for all $T = 1,\ldots, n$ [15]. The piecewise function is needed to provide flexibility in the model in terms of modelling the outbreaks of dengue fever in addition to possible seasonal trends that we observe.

The two-component model formulation is completed by specifying prior distributions for the parameters in the model as follows:

$$\begin{aligned} \boldsymbol{\gamma} &\sim N\left(0, \sigma_\gamma^2 \mathbf{I}\right), \\ \lambda &\sim \mathrm{Ga}\left(\chi_\xi, \delta_\xi\right), \\ \psi &\sim \mathrm{S}\left(\alpha_\psi, \beta_\psi\right) \end{aligned} \tag{5}$$

N denotes a normal and Ga a Gamma distribution. $\sigma_\gamma{}^2$ was set to 10^6, representing highly dispersed independent normal priors for each coefficient. **I** is an identity matrix. For $\lambda^{(k)}$, $k = 1,\ldots, K + 1$, independent exponential distributions with mean $1/\xi$ and variance $1/\xi^2$ were specified. ξ was then assigned a gamma hyperprior Ga(χ_ξ, δ_ξ). The marginal prior distribution for $\lambda^{(k)}$ is then a gamma-gamma distribution [22]. In our study, we used $\chi_\xi = 10$ and $\delta_\xi = 10$, which corresponded to the gamma-gamma marginal of $\lambda^{(k)}$ turning out to be an F-distribution with (2,20) degrees of freedom, which then indicates that the marginal prior probability of an outbreak occurring (i.e., $\lambda^{(k)} \geq 1$) is 0.39, while always favouring smaller values of $\lambda^{(k)}$, with the density function monotonically decreasing. The dispersion parameter for the negative binomial distribution, ψ, which was designed to handle extra-Poisson variation in the data, was assigned a gamma hyperprior as well, with the following parameter, Ga(α_ψ, β_ψ). α_ψ and β_ψ were assigned values of 1 and 0.1, respectively in the original analysis corresponding to a prior mean and standard deviation of 10.

The K-H models were fitted using the customised Bayesian software Twins V1.0 [15]. Markov Chain Monte Carlo (MCMC) methods, in particular the Metropolis-Hastings algorithm, were used to estimate the parameters. For each model, we ran 200 iterations as burn-ins. These burn-in samples were discarded and not used in the analysis. We ran a further 60,000 iterations, but only saved every 20th observation, resulting in a final 3000 sample size. This was to circumvent the problem of autocorrelated samples.

2.4. Model Comparison

We compared the ARIMA model with the K-H model and as well conducted a sensitivity analysis on the K-H model using the MAPE:

$$\text{MAPE} = \frac{1}{n}\sum_{T=1}^{n}\left|\frac{f_T - F_T}{F_T}\right|, \tag{6}$$

where n is the total number of weeks of data.

The Bayesian analyses were based on several assumptions regarding the prior distributions, and we assessed the robustness of our results in a sensitivity analysis. For the sensitivity analyses, we considered 4 different scenarios which involved varying values of χ_ξ, δ_ξ or α_ψ, β_ψ while keeping the other variables at their original values: Model 1: $\alpha_\psi = 0.1$ and $\beta_\psi = 0.1$, Model 2: $\alpha_\psi = 10$ and $\beta_\psi = 1$, Model 3: $\chi_\xi = 1$ and $\delta_\xi = 1$, and Model 4: $\chi_\xi = 10$ and $\delta_\xi = 1$. The prior values for the sensitivity analysis were selected to represent a range of realistic scenarios where the probabilities of an outbreak were expected to be different. In particular, we selected priors where the probability of observing an outbreak ranged from 0.001 (for $\chi_{\xi} = 10$ and $\delta_{\xi} = 1$) to 0.5 ($\chi_{\xi} = 1$ and $\delta_{\xi} = 1$).

3. Results

Go to: Go to:

Figure 1 highlights the weekly distribution of DF notifications in Singapore from January 2001 to June 2008. It is evident that DF notification exhibits both seasonal trends (e.g., regular peaks around June or September and troughs seen in the first 4 months of the year) and epidemic trends (most markedly shown during the 2005 epidemic, when average weekly counts exceeded 600 cases).

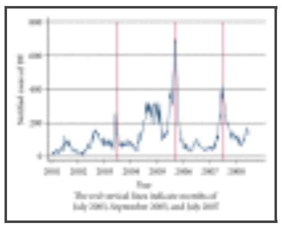


Figure 1
Weekly cases of dengue fever (DF) in Singapore.

The autocorrelation plots for DF (Figure 2(a)) indicated that correlations gradually declined over the weeks to insignificant values after 12 weeks. The partial autocorrelations plots (Figure 2(b)) showed a spike at week 1 and week 4 indicating possible inclusion of AR terms of the order of up to four in the ARIMA model. We evaluated the various combinations, including autocorrelation terms 3 and 4 in our analysis.

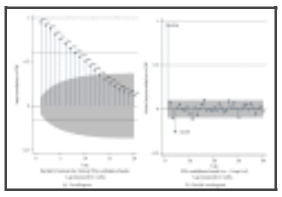


Figure 2
Plots of autocorrelation and partial correlation for dengue fever (DF).

We explored various formulations of the ARIMA model, and we summarise some of the more important ones in Table 1. As can be seen, ARIMA (3,1,0) provided the lowest MAPE value of 19.86. Including a moving average term did not improve the fit of the model, as with adding an autocorrelation term of four. Adding a 12-month seasonal component (not shown) also did not lower the MAPE. The parameters for the final ARIMA model are shown in Table 2. We found all three autoregressive terms AR(1) = −0.10 ($P = 0.001$), AR(2) = 0.10 ($P = 0.002$), and AR(3) = 0.23 ($P <$ 0.001) to be statistically significant. The parameters for the K-H model are also provided in Table 2.

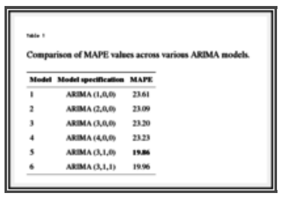


Table 1

Comparison of MAPE values across various ARIMA models.

Model	Model specification	MAPE
1	ARIMA (1,0,0)	23.61
2	ARIMA (2,0,0)	23.09
3	ARIMA (3,0,0)	23.20
4	ARIMA (4,0,0)	23.23
5	ARIMA (3,1,0)	**19.86**
6	ARIMA (3,1,1)	19.96

Table 1
Comparison of MAPE values across various ARIMA models.

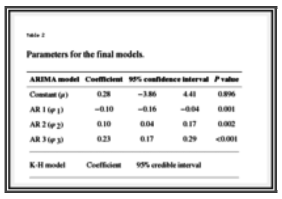


Table 2

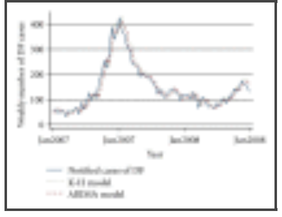
Parameters for the final models.

ARIMA model	Coefficient	95% confidence interval		P value
Constant (μ)	0.28	−3.86	4.41	0.896
AR 1 (φ_1)	−0.10	−0.16	−0.04	0.001
AR 2 (φ_2)	0.10	0.04	0.17	0.002
AR 3 (φ_3)	0.23	0.17	0.29	<0.001
K-H model	Coefficient	95% credible interval		

Table 2
Parameters for the final models.

The comparison between the ARIMA and K-H model is shown in Figure 3. Table 3 shows the results from comparing the two models. Overall, the K-H model performed marginally better than the ARIMA model (MAPE of 17.21 and 17.54 resp.). In particular, the model predicted well (out-of-sample) for certain periods, including the early endemic periods between weeks 1 to 12. Fine-tuning the parameters for the K-H model allowed us to make better predictions for the epidemic periods, as we show in the sensitivity analysis (Table 4). For instance, the model predicted well for the epidemic periods within the weeks 17 to 24 (sensitivity analysis 4).

Figure 3
Comparison of out-of-sample forecasts of dengue fever (DF) betwe



two-component K-H model (January 2007 to June 2008).

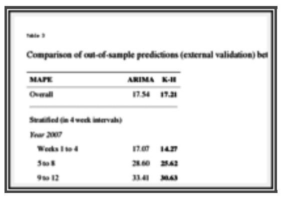


Table 3

Comparison of out-of-sample predictions (external validation) bet

MAPE	ARIMA	K-H
Overall	17.54	**17.21**
Stratified (in 4 week intervals)		
Year 2007		
Weeks 1 to 4	17.07	**14.27**
5 to 8	28.60	**25.62**
9 to 12	33.41	**30.63**

[Table 3](#)
Comparison of out-of-sample predictions (external validation) between ARIMA and K-H models.

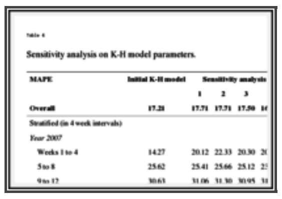


Table 4

Sensitivity analysis on K-H model parameters.

MAPE	Initial K-H model	Sensitivity analysis		
		1	2	3
Overall	17.21	17.71	17.71	17.50
Stratified (in 4 week intervals)				
Year 2007				
Weeks 1 to 4	14.27	20.12	22.33	20.30
5 to 8	25.62	25.41	25.66	25.12
9 to 12	30.63	31.06	31.30	30.95

Table 4
Sensitivity analysis on K-H model parameters.

In terms of forecasting one-week ahead DF notifications, both methods performed well (Figure 3). For instance, the K-H model forecasted 53 (observed 58) for 2007 week 1, 356 (observed 371) for 2007 week 26, 112 (observed 115) for 2008 week 1, and 171 (observed 132) for 2008 week 26. It is worth noting that these results are for out-of-sample predictions.

The Bayesian analysis is influenced by the prior specification. As such, we investigated the robustness of our results to different formulation of the priors. These priors represented a wide range of realistic scenarios where the probability of an outbreak is expected to differ. As can be seen from Table 4, it appears that the models have generally similar MAPE values except for sensitivity analysis 4, where the MAPE is actually the lowest at 16.54. In our local setting, we found that specifying a small prior probability of 0.001 for an outbreak to occur provided a better fit of the data.

4. Discussion

Go to: Go to:

We found that the K-H model performed better than the conventional ARIMA time series model; however, this was only marginal. Forecasting weekly cases of DF has immense implication for hospital resources planning. For an infectious disease ward, knowing the normal trend of DF, along with predictions of the following week's DF can allow hospital planners to better plan for and allocate their manpower and other resources. Intensive media campaigns (e.g., television advertisements) in the weeks prior to a projected increase in DF notifications may prove to reduce the number of new cases.

Though we used the MAPE index to compare the models, other indices are also available. The Mean Squared Error, for instance, is calculated from the sum of the squared error values. Compared to MAPE, the values are not relative to the magnitude of the observation, and the values are not intuitively easy to interpret.

There were several limitations in our study. Firstly, our analysis was dependent on notifiable data. While clinicians are required to report all cases of DF and DHF to the MOH, there is a possibility that the cases could be underreported, especially since mild asymptomatic cases of DF may have not been diagnosed. While this may have led to an under-estimate in the forecasts, the comparisons across the models are still valid, as they make use of the same number of weekly cases.

In our analysis, we compared the predictive capability of the models using one-week ahead forecast of dengue fever notification. It is possible to forecast for periods longer than that, of course the predictions may inherently not be as accurate as a one-week forecast.

In conclusion, we found that both the final models chosen for the ARIMA and K-H models predict the future course of DF in Singapore reliably well, while the former performed marginally better. The ARIMA models were relatively faster to implement and run, while the K-H model was sensitive to the choice of priors, which needs to be carefully made before the study is conducted.

Acknowledgments

Go to: Go to:

Funding for this study was received from the Duke-NUS SRP block grant as well as the Merit Award from the Yong Loo Lin School of Medicine, Singapore.

References

Go to: Go to:

1. World Health Organisation. *Revised March*. 117. Geneva, Switzerland: World Health Organisation; 2009. Dengue and dengue haemorrhagic fever.

2. Ooi EE, Goh KT, Gubler DJ. Dengue prevention and 35 years of vector control in Singapore. *Emerging Infectious Diseases*. 2006;12(6):887–893. [PMC free article] [PubMed]

3. Chowell G, Torre CA, Munayco-Escate C, et al. Spatial and temporal dynamics of dengue fever in Peru: 1994–2006. *Epidemiology and Infection*. 2008;136(12):1667–1677. [PMC free article] [PubMed]

4. Mammen MP, Pimgate C, Koenraadt CJM, et al. Spatial and temporal clustering of dengue virus transmission in Thai villages. *PLoS Medicine*. 2008;5(11, article no. e205):1605–1616. [PMC free article] [PubMed]

5. Mondini A, Chiaravalloti-Neto F. Spatial correlation of incidence of dengue with socioeconomic, demographic and environmental variables in a Brazilian city. *Science of the Total Environment*. 2008;393(2-3):241–248. [PubMed]

6. Koh BKW, Lee CN, Kita Y, et al. The 2005 dengue epidemic in Singapore: epidemiology, prevention and control. *Annals of the Academy of Medicine Singapore*. 2008;37(7):538–545. [PubMed]

7. Brunkard JM, Cifuentes E, Rothenberg SJ. Assessing the roles of temperature, precipitation, and ENSO in dengue re-emergence on the Texas-Mexico border region. *Salud Publica de Mexico*. 2008;50(3):227–234. [PubMed]

8. Chadee DD, Shivnauth B, Rawlins SC, Chen AA. Climate, mosquito indices and the epidemiology of dengue fever in Trinidad (2002-2004) *Annals of Tropical Medicine and Parasitology*. 2007;101(1):69–77. [PubMed]

9. Depradine CA, Lovell EH. Climatological variables and the incidence of Dengue fever in Barbados. *International Journal of Environmental Health Research*. 2004;14(6):429–441. [PubMed]

10. Eearnest A, Tan SB, Wilder-Smith A. Meteorological factors and El Nino Southern Oscillation are independently associated with dengue infections. *Epidemiology and Infection*. 2011;12:1–8. [PubMed]

11. Moineddin R, Upshur REG, Crighton E, Mamdani M. Autoregression as a means of assessing the strenght of seasonality in a time series. *Population Health Metrics*. 2003;1, article no. 10 [PMC free article] [PubMed]

12. Reis BY, Mandl KD. Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making*. 2003;3(1):p. 2. [PMC free article] [PubMed]

13. Hay SI, Myers MF, Burke DS, et al. Etiology of interepidemic periods of mosquito-borne disease. *Proceedings of the National Academy of Sciences of the United States of America*. 2000;97(16):9335–9339. [PMC free article] [PubMed]

14. Heisterkamp SH, Dekkers ALM, Heijne JCM. Automated detection of infectious disease outbreaks: hierarchical time series models. *Statistics in Medicine*. 2006;25(24):4179–4196. [PubMed]

15. Held L, Hofmann M, Höhle M, Schmid V. A two-component model for counts of infectious diseases. *Biostatistics*. 2006;7(3):422–437. [PubMed]

16. O’Neill PD. A tutorial introduction to Bayesian inference for stochastic epidemic models using

Markov chain Monte Carlo methods. *Mathematical Biosciences*. 2002;180:103–114. [PubMed]

17. Hay JL, Pettitt AN. Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease. *Biostatistics*. 2001;2(4):433–444. [PubMed]

18. He D, Ionides EL, King AA. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface*. 2010;7(43):271–283. [PMC free article] [PubMed]

19. King AA, Ionides EL, Pascual M, Bouma MJ. Inapparent infections and cholera dynamics. *Nature*. 2008;454(7206):877–880. [PubMed]

20. Infectious Diseases Guidelines. http://www.moh.gov.sg/content/moh_web/home/Publications/guidelines/infectious_diseases_guidelines/2011/a_guide_on_infectious_diseases_of_public_health_importance_in_singapore.html, 2011.

21. Stata V11.0. Stata Corp College Station, Texas, USA.

22. Bernado JM, Smith AFM. *Bayesian Theory*. London, UK: Chapman & Hall; 1994.