J Vector Borne Dis 51, September 2014, pp. 172–178

Forecasting Japanese encephalitis incidence from historical morbidity patterns: Statistical analysis with 27 years of observation in Assam, India

Bijoy K. Handique[1], Siraj A. Khan[2], J. Mahanta[2] & S. Sudhakar[1]

[1]*North Eastern Space Applications Centre, Umiam, Meghalaya;* [2]*Regional Medical Research Centre-NER (ICMR), Dibrugarh, Assam, India*

ABSTRACT

Background & objectives: Japanese encephalitis (JE) is one of the dreaded mosquito-borne viral diseases mostly prevalent in south Asian countries including India. Early warning of the disease in terms of disease intensity is crucial for taking adequate and appropriate intervention measures. The present study was carried out in Dibrugarh district in the state of Assam located in the northeastern region of India to assess the accuracy of selected forecasting methods based on historical morbidity patterns of JE incidence during the past 22 years (1985–2006).

Methods: Four selected forecasting methods, *viz.* seasonal average (SA), seasonal adjustment with last three observations (SAT), modified method adjusting long-term and cyclic trend (MSAT), and autoregressive integrated moving average (ARIMA) have been employed to assess the accuracy of each of the forecasting methods. The forecasting methods were validated for five consecutive years from 2007–2012 and accuracy of each method has been assessed.

Results: The forecasting method utilising seasonal adjustment with long-term and cyclic trend emerged as best forecasting method among the four selected forecasting methods and outperformed the even statistically more advanced ARIMA method. Peak of the disease incidence could effectively be predicted with all the methods, but there are significant variations in magnitude of forecast errors among the selected methods. As expected, variation in forecasts at primary health centre (PHC) level is wide as compared to that of district level forecasts.

Interpretation & conclusion: The study showed that adopted forecasting techniques could reasonably forecast the intensity of JE cases at PHC level without considering the external variables. The results indicate that the understanding of long-term and cyclic trend of the disease intensity will improve the accuracy of the forecasts, but there is a need for making the forecast models more robust to explain sudden variation in the disease intensity with detail analysis of parasite and host population dynamics.

Key words Forecasting; Japanese encephalitis; morbidity pattern; time-series analysis

INTRODUCTION

Japanese encephalitis (JE) is a dreaded vector (mosquito) borne viral disease mostly prevalent in Asian countries including India. Virus from wild birds through vector mosquitoes spreads to peridomestic and domestic birds and then to mammals like cattle, pigs, *etc.* and eventually spills over to man. JE virus has been isolated from about 30 species of mosquitoes worldwide which fall under five genera, *viz. Culex, Anopheles, Aedes, Armigeres* and *Mansonia*. However, only a few species meet the requirements to be classified as important vectors[1]. Since, the first record of JE case in India in 1955 in Tamil Nadu followed by isolation of JE virus from wild caught mosquitoes in 1956, in the last couple of decades, epidemics of JE have occurred in the states of West Bengal, Assam, Manipur, Nagaland, Uttar Pradesh, Bihar and Goa in addition to south India[2–3]. JE cases have attained alarming proportions to pose a major public health problem in India, more so due to unavailability of any cure for the disease and due to its quite high case: fatality ratio[4].

Existing global systems for epidemic preparedness focus on disease surveillance using either expert knowledge or statistical modeling of disease activity and thresholds to identify time and areas of risk[5]. Geospatial techniques comprising of remote sensing (RS), geographical information system (GIS) and global positioning system (GPS) have also emerged as effective tools for surveillance of vector habitats and risk assessment[6–10]. In addition, spatial analysis and geostatistical analysis tools have helped in integrating wide range of attribute parameters and building effective models for disease forewarning[11–14]. The potential use of time series techniques in epidemiological studies, disease surveillance and outbreak forecast has been explored in many studies[15–19].

Considering the large geographical coverage and local health care systems in a country like India, developing an early warning system at a district level has been

imperative. Accurate disease forecasting models would markedly improve epidemic prevention and control capabilities. Specific forecasts of incidence would be helpful to local health services for appropriate preparedness and to take selective preventive measures in areas at risk of epidemics[20].

In this study, we explored the possibility to forecast JE incidence from the patterns of historical morbidity data alone (without external predictors) while making use of the data on disease intensity and spatial distributions at the primary health centre (PHC) level. Different forecasting methods have been employed to assess the accuracy of each of the forecasting methods. The forecasting methods were validated for five consecutive years from 2007–2012 and accuracy of each method was determined by calculating errors resulting from the difference between the observed and forecasted JE incidence.

MATERIAL & METHODS

Study area

The study was carried out in Dibrugarh district of Assam state located in the northeastern part of India considering the severity of impact of JE and its perennial occurrence (average annual case load in Assam during the last two decades, since 1980 has been 295 and the average annual incidence per million population being 12.5). Dibrugarh district alone shared a burden of 37 cases per million population in the area. The district covering a geographical area of 7023.9 km^2 lies between 27° 15′ N – 28° N latitude and 94° 45′ E – 96° E longitude (Fig. 1). The district is divided into six PHCs, *viz.* Barbaruah, Lahowal, Panitola, Tengakhat, Khowang and Naharani for monitoring and providing health care services in the district.

Information flow of JE cases

Assam Medical College Hospital (AMCH) located in the district headquarter of Dibrugarh is the only specialised hospital for JE treatment in the district. Suspected AES (acute encephalitis syndrome) reported at different PHCs or at private hospitals are referred to the AMCH. Blood samples from the AMCH along with patient records are sent to the Regional Medical Research Centre-NE Region (RMRC–ICMR) located in Dibrugarh for laboratory confirmation of JE cases. Details of JE cases with the address of the patients are recorded by AMCH record department, which are also sent to Joint Director of Health Services at Dibrugarh and State Directorate of Health Services in Guwahati. We collected the JE case data from RMRC, Dibrugarh, cross matched with the data available in the records department of AMCH, as well as from the Joint Director of Health Services in Dibrugarh, thus minimising chances of missing any information.

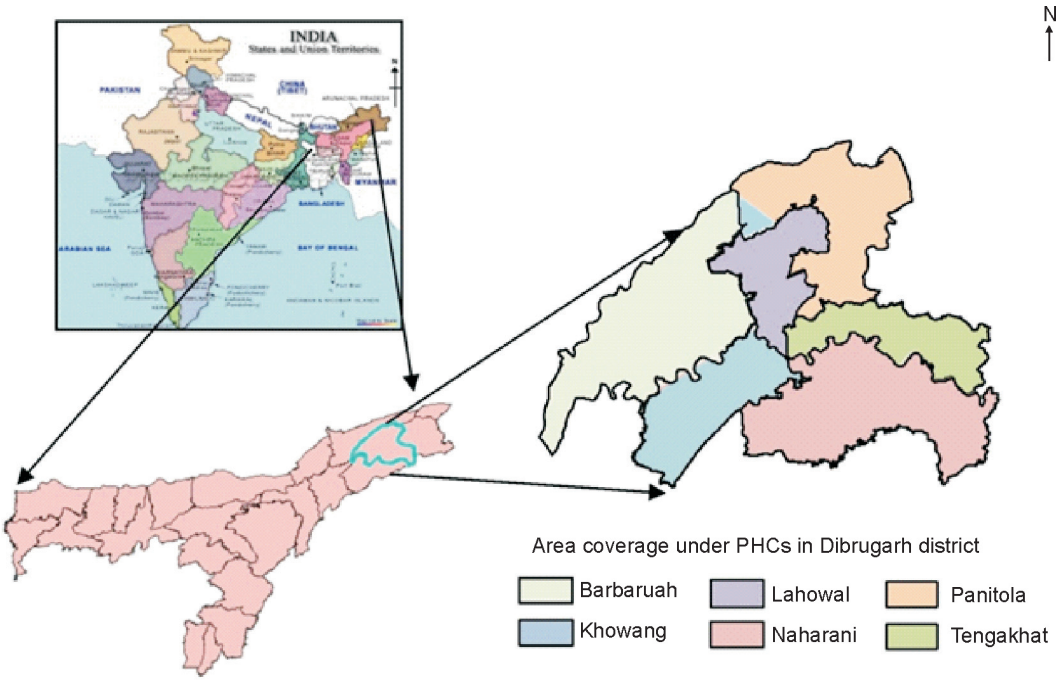

Fig. 1: Location map of the study area.

PHC wise forecasting of JE cases

PHC wise JE incidence from 1985–2006 has been observed to follow approximately a log normal distribution. So, it has been decided to carry out the analysis based on log-transformed series in the similar method adopted by Abeku *et al*[21]. Relative incidence (RI) of the disease has been calculated to bring the data on disease incidence collected from different areas into the same scale. The RI from the month t (denoted by Y_t) is calculated as:

$$Y_t = \frac{\ln Z_t}{A}$$

where, Z_t is the number of cases in month t and A is the overall mean of the log-transformed series used for forecasts. The back-transformed number of cases is thus:

$$Z_t = \exp(AY_t)$$

The mean (A) differs for each series or sample.

Forecasting methods

The following methods were used to forecast RI m months in advance, i.e. to obtain the forecast for the month $t+m$ denoted as $\hat{Y}_{t+m}$. These methods were compared in terms of their forecast accuracy and discussed below in order of their complexity.

Seasonal average (SA): This method uses the historical average of each particular calendar month as forecast for the same month in the future, which means the average of all observed RI values during the same calendar month in previous years will be the forecast value for the corresponding month in the future.

$$\hat{Y}_{t+m} = A_{t+m}$$

Seasonal adjustment with last three observations (SAT): The seasonal average was corrected using the mean deviation of three most recent observations from their expected seasonal values to generate forecasts for future months. The object was to capture trend in incidence during the most recent months while reducing statistical variation:

$$\hat{Y}_{t+m} = A_{t+m} + \frac{\sum_{i=0}^{2}(Y_{t-i} - A_{t-i})}{3}$$

where, $t–i$ denotes a month i lags before the (last) month t.

Modified method adjusting long-term and cyclic trend (MSAT): A modification to the above method is applied to adjust the cyclic and long-term trend on JE intensity in a particular month. A best fitted trend curve selected to determine a forecast value of one year forward. A three yearly moving average has been applied to adjust the cyclic trend in JE intensity. Out of these two trend values, closer point to Y_t is selected and denoted by T_m. The back-transformation to the average value of $\hat{Y}_{t+m}$ and T_m will give the predicted value of JE cases.

$$\hat{Y}_{t+m} = \frac{1}{2}\left[\left(A_{t+m} + \frac{\sum_{i=1}^{2}\left(Y_{t-i} - A_{t-i}\right)}{3}\right) + T_m\right]$$

Autoregressive integrated moving average (ARIMA): The autocorrelation pattern in each series at different lags was used to develop ARIMA models[22–23]. A single equation ARIMA model states how any value in a single time series is linearly related to its own past values through combining two processes: the autoregressive (AR) process which expresses Y_t as a function of its past values, and the moving average (MA) process, which expresses Y_t as a function of past values of the error term e as—

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + - - + \alpha_p Y_{t-p} - e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - - - - \beta_q e_{t-q}$$

where, the α and β are the coefficients of AR and MR processes respectively and p and q are the number of past values of Y_t and the error term used respectively. Application of ARIMA technique requires the series to be stationary, i.e. constant mean and variance over time. A series with constant variance can be obtained by applying log and other type of transformation to the original series. A constant mean can be obtained by taking the first or higher order difference of the variable as necessary until the series become stationary.

Accuracy of forecast

It is important to know the accuracy of a forecast. Theil's coefficient is a reliable measure to judge the accuracy of forecast[24–25]. Theil's coefficient is given by:

$$U = \frac{\sqrt{\frac{1}{n}\sum(F_i - A_i)^2}}{\sqrt{\frac{1}{n}\sum F_i^2} - \sqrt{\frac{1}{n}\sum A_i^2}}$$

where, F_i, the series represents the forecasted values of a time series and A_i is the real value of the same series. Value of 'U' lies between 0 and 1. 'U' value near to 0 forecasting is considered to be accurate.

Whether the relation between forecasted and observed

disease cases measured in terms of Pearson's coefficient (r) is statistically significant was tested with the help of t-test, where t is given by:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{(n-2)} \text{ with (n -2) degrees of freedom.}$$

Different analyses in GIS domain have been performed using ARC GIS 9.3 Software[26]. Statistical analyses have been done using Microsoft Excel and SPSS 19.0.

RESULTS & DISCUSSION

JE transmission in almost all the PHCs was observed to be highly variable from season to season and year to year, although the month of July has been observed to be the peak for all the six PHCs during the years from 2007–2012. This may be due to the fact that infectious diseases particularly those transmitted by intermediate hosts are known to be highly sensitive to long-term changes in climate and short-term fluctuations in the weather[27]. Based on this long-term and short-term fluctuations in the disease occurrence ARIMA and SARIMA (seasonal ARIMA) models have been widely used for epidemic time series forecasting including the hemorrhagic fever with renal syndrome[28–29], dengue fever[30–31], and tuberculosis[32]. Moreover, as there have been many different time series models for prediction, it will be of genuine interest to evaluate the best suitability of a model for the prediction of epidemic incidence. Where comparative studies on the accuracy of different models for forecasting epidemic behaviour were carried out, inconsistency in model performance between studies has been observed. For example, SARIMA model had demonstrated better performance than generalized models in forecasting cryptosporidiosis cases in northeastern Spain[33], and better than regression and decomposition models in forecasting campylobacteriosis in the United States of America[34], but dynamic linear models showed better performance than the SARIMA model in forecasting hepatitis-A and malaria[35]. The different findings of these studies suggest that further studies focusing on the comparison of different kinds of predicting methods for different types of diseases are necessary for the application in forecasting epidemic behaviour.

Highest forecast accuracy was observed with the modified method with seasonal adjustment adjusting long-term cyclic trend (MSAT) followed by the method of seasonal adjustment with last three observations. It has been noted that the relation between forecasted and observed disease cases by MSAT measured in terms of Pearson's coefficient (r) has been found to be statistically significant ($p<0.05$) tested with t-test.

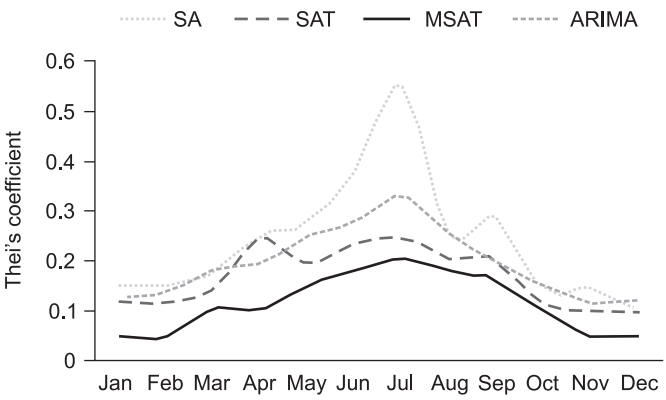

Fig. 2: Accuracy of forecast with different forecasting techniques.

ARIMA method resulted in better forecast than the seasonal average. This implies the highest forecast error in case of SA followed by ARIMA and SAT methods during the validation period from 2007–2012 (Fig. 2). Other methods such as INAR modeling or Markov chain analysis, which can be applied to situations like this, where ARIMA modeling fails, but they are less practical[36].

Considering the fact that MSAT technique provided better forecast, we have validated the forecast for five years at PHC level. Figures 3–8 depict the forecasted and actual number of cases during 2007–2012. It has been observed that in case of sudden rise in the JE cases, the forecast model could reasonably predict a higher intensity but in no case it exactly reached the peak. Considering that there is rise in JE cases in the study areas during recent years, the adjustment in last three years of observation could adequately explain the variations.

It is interesting to note that there is less variation at the district level for the forecasts made with selected forecasting techniques. Due to adjustment in the long-term and cyclic trend of the disease, MSAT method resulted in a better forecast accuracy at the district level. As expected SA method resulted in the highest forecast error for the validating years. It has also been observed that due to equal weightage of morbidity pattern of all the earlier years, SA method resulted in lower predicted values through all the years. Actual and forecasted JE cases with selected forecasting methods in Dibrugarh district during 2007–2012 are depicted in Fig. 9.

Another important observation was that inspite of statistical sophistications, ARIMA method could not yield expected forecast accuracy either at the PHC level or at the district level. Other studies have also indicated that the statistically advanced ARIMA models may produce

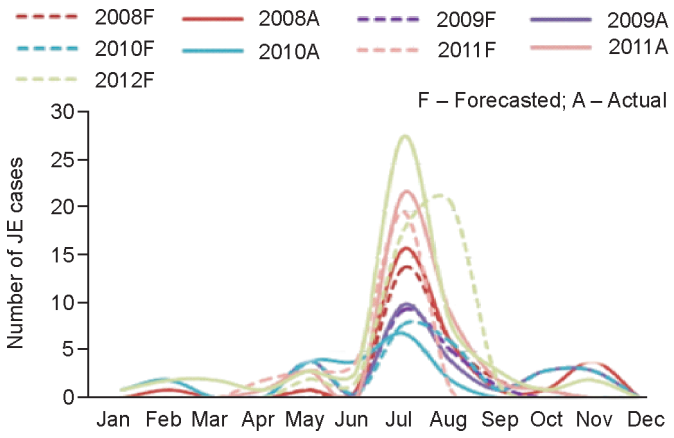

Fig. 3: Forecasted and actual number of JE cases in Barbaruah PHC.

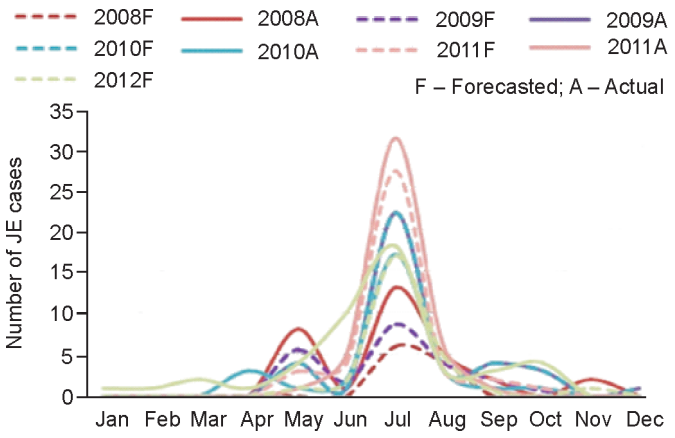

Fig. 6: Forecasted and actual number of JE cases in Naharani PHC.

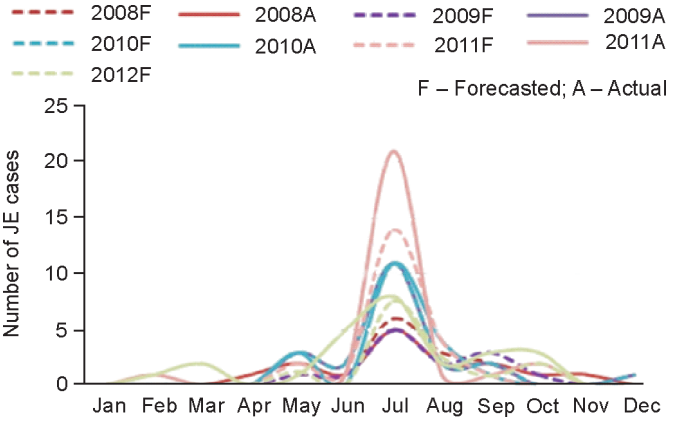

Fig. 4: Forecasted and actual number of JE cases in Khowang PHC.

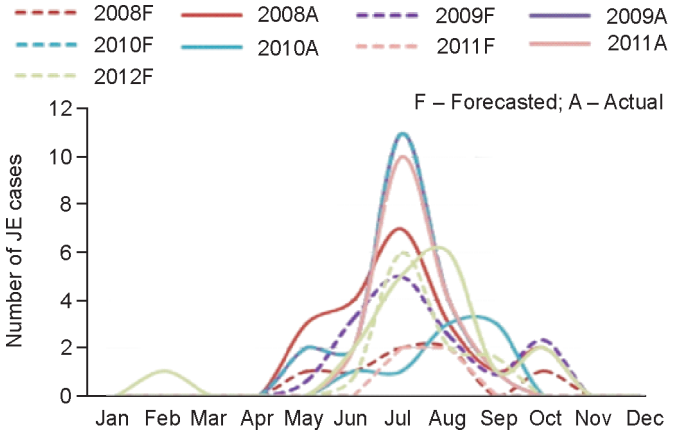

Fig. 7: Forecasted and actual number of JE cases in Panitola PHC.

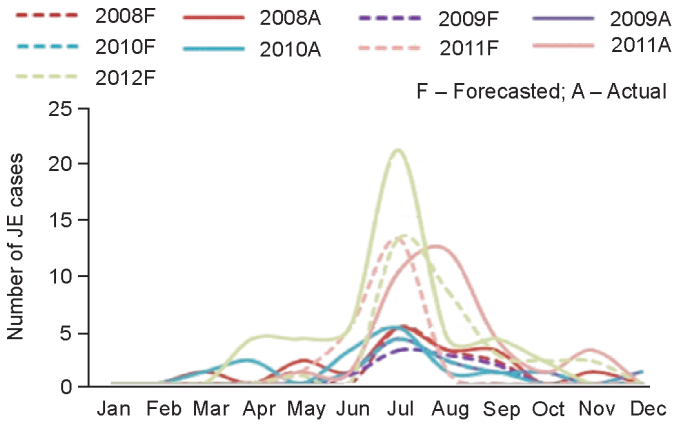

Fig. 5: Forecasted and actual number of JE cases in Lahowal PHC.

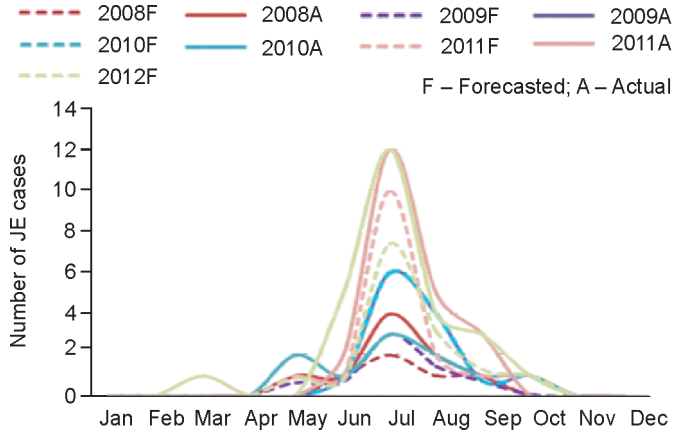

Fig. 8: Forecasted and actual number of JE cases in Tengakhat PHC.

very good fit to the data but in post-sample forecast, they would not be robust enough to handle a possible change in behaviour of the series[37].

Spatial models are being used with increasing frequency to help characterize these large-scale patterns and to evaluate the impact of interventions. It also demonstrates the need to develop a simple model of household demographics, so that large-scale models can be extended to the investigation of long-time scale human pathogens[38]. But collection of adequate and enough data has always been an issue of concern[39].

CONCLUSION

Progress in mathematical analysis and modeling is of fundamental importance to our growing understanding

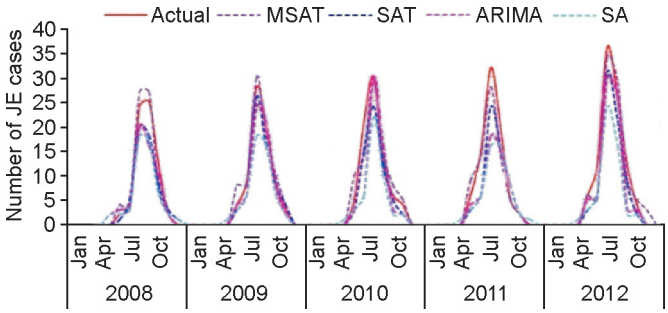

Fig. 9: Forecasted and actual JE cases in Dibrugarh district during 2007–2012.

of pathogen evolution and ecology. The fit of mathematical models to surveillance data has informed both scientific research and health policy[40].

The study shows that adopted forecasting techniques could reasonably forecast the intensity of JE cases at PHC level without considering the external variables. Such quick forecasts will be of immense help for health authorities to prepare for intervention measures that do not require complex analysis of disease vector and host dynamics. However, it may be of interest to assess whether a combination of seasonal climate forecasts, monitoring of meteorological conditions, and early detection of cases could scale down the emergency[41]. Results shown by MSAT technique indicate that appropriate understanding of long-term and cyclic trend of the disease intensity will improve the accuracy of the forecasts. It has also been noted that even the MSAT method is not robust enough to explain sudden variation in the disease intensity, and may require an approach that uses a detail analysis of parasite and host population dynamics, control measures adopted, *etc.* In addition, the application of environmental data to the study of disease will offer the capability to demonstrate vector-environment relationships and potentially forecast the risk of disease outbreaks or epidemics.

ACKNOWLEDGEMENT

The authors would like to thank Co-Scientists of North-Eastern Space Applications Centre, Umiam and Regional Medical Research Centre-NER (ICMR), Dibrugarh for their encouragement. Sincere thanks are also due to Joint Director of Health Services, Dibrugarh district and authorities of Assam Medical College, Dibrugarh for providing data on JE cases in the study area. Financial support received from Department of Space, Govt. of India for the study is duly acknowledged.

REFERENCES

1. Vythilingam I, Mahadevan S, Tan SK, Abdulla G, Jeffery J. Host feeding pattern and resting habits of Japanese encephalitis vectors found in Sepang, Selangor, Malaysia. *Trop Biomed* 1996; *13:* 45–50.
2. Kabilan L, Rajendran R, Arunachalam N, Ramesh S, Srinivasan S, Philip S, *et al*. Japanese encephalitis in India: An overview. *The Indian J Paediatr* 2004; *71*(7): 609–15.
3. Khan SA, Narain K, Handique R, Dutta P, Mahanta J, Satyanarayana K, *et al*. Role of some environmental factors in modulating seasonal abundance of potential Japanese encephalitis vectors in Assam. *J Trop Med Public Health* 1996; *27* (2): 382–91.
4. Phukan AC, Baruah PK, Mahanta J. Japanese encephalitis in Assam, North East India. *Southeast Asian J Trop Med Public Health* 2004; *35:* 618–22.
5. Myers MF, Rogers DJ, Cox J, Flahault A, Hay SI. Forecasting disease risk for increased epidemic preparedness in public health. *Adv Parasitol* 2000; *47:* 309–30.
6. Abelardo C, Moncayo John DE, John TF. Application of geographic information technology in determining risk at Eastern Equine Encephalomyelitis virus transmission. *J Am Mosq Control Assoc* 2000; *16*(1): 28–35.
7. Beck EL, Roddriguez MH, Dister SW, Rodriguez AD, Rejmankova E, Ulloa A, *et al*. Remote sensing as a landscape epidemiological tool to identify village at high risk for malaria transmission. *Am J Trop Med Hyg* 1994; *51*(3): 271–80.
8. Dhiman RC. Remote sensing: A visionary tool in malaria epidemiology. *ICMR Bull* 2000; *30*(11): 1–2.
9. Glass GE, Schwartz BS, Morgan JM, Johnson DT, Noy PM, Israel E. Environmental risk factor for lyme disease identified with Geographic Information System. *Am J Public Health* 1995; *85:* 944–8.
10. Hugh Jones M. Application of remote sensing to the identification of the habitats of parasite and disease vectors. *Parasitol Today* 1989; *5*(8): 244–51.
11. Gaetan C, Guyon X. *Spatial statistics and modeling.* New York, USA: Springer 2010; p. 25–47.
12. Chou YH. *Exploring spatial analysis in Geographic Information Systems*. Santa Fe, USA: Onword Press 1997; p. 202–5.
13. Lawson AB. *Statistical methods in spatial epidemiology.* West Sussex, England: John Wiley and Sons Ltd 2001; p. 3–25.
14. Sabesan S, Konuganti HKR, Perumal V. Spatial delimitation, forecasting and control of Japanese encephalitis: India –A case study. *Open Parasitol J* 2008; *2:* 59–63.
15. Shumway RH, Stoffer DS. *Time series analysis and its applications with R examples.* III edn. London: Springer 2011; p. 83–154.
16. Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White Nicholas J, Kaewkungwal J. Development of temporal modeling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhutan. *Malar J* 2010; *9:* 251.
17. Allard R. Use of time-series analysis in infectious disease surveillance. *Bull World Health Organ* 1998; *76:* 327–33.
18. Helfenstein U. The use of transfer function models, intervention analysis and related time-series methods in epidemiology. *Int J Epidemiol* 1991; *20:* 808–15.
19. Hay SI, Were EC, Reneshaw M, Noor AM, Ochaola SA, Olusanmi L, *et al*. Forecasting, warning, and detection of malaria epidemics: A case study. *Lancet* 2003; *361:* 1705–6.

20. Hay SI, Rogers DJ, Shanks DG, Myers MF, Snow RW. Malaria early warning in Kenya. *Trends Parasitol* 2001; *17:* 95–9.
21. Abeku TA, de Tarekegn A, Vlas SJ, Borsboom G, Teklehaimanot A, Kebede A, *et al*. Forecasting malaria incidence from historical morbidity patterns in epidemic-prone areas of Ethiopia: A simple seasonal adjustment method performs best. *Trop Med Int Health* 2002; *7*(10): 851–7.
22. Box GEP, Jenkins GM. *Time series analysis: Forecasting and control.* Rev edn. San Francisco: Holden Day 1976; p. 41–50.
23. Brockwell PJ, Davis RA. *Time series: Theory and methods.* II edn. New York: Springer-Verlag 1991; p. 77–91.
24. Bliemel F. Theil's. Forecast accuracy coefficient— A clarification. *J Marketing Res* 1973; *10* (4): 444–6.
25. Armstrong JS. *Principles of forecasting— A handbook for researchers and practitioners.* London: Springer 2001; p. 443–70.
26. Lee J, Wong DWS. *Statistical analysis with ARCVIEW GIS.* New York: John Wiley and Sons 2001; p. 135–89.
27. Myers MF, Rogers DJ, Cox J, Flahault A, Hay SI. Forecasting disease risk for increased epidemic preparedness in public health. *Adv Parasitol* 2000; *47:* 309–30.
28. Li Q, Guo N-N, Han Z-Y, Zhang Y-B, Qi S-X, Xu YG. Application of an autoregressive integrated moving average model for predicting the incidence of hemorrhagic fever with renal syndrome. *Am J Trop Med Hyg* 2012; *87:* 364–70.
29. Liu Q, Liu X, Jiang B, Yang W. Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model. *BMC Infect Dis* 2011; 11: 218. 10.1186/1471-2334-11–218.
30. Luz PM, Mendes BVM, Codeco CT, Struchiner CJ, Galvani AP. Time series analysis of dengue incidence in Rio de Janeiro, Brazil. *Am J Trop Med Hyg* 2008; *79:* 933–9.
31. Wongkoon S, Jaroensutasinee M, Jaroensutasinee K. Development of temporal modeling for prediction of dengue infection in Northeastern Thailand. *Asian Pacific J Trop Med* 2012; *5:* 249–52.
32. Rios M, Garcia JM, Sanchez JA, Perez DA. Statistical analysis of the seasonality in pulmonary tuberculosis. *European J Epidemiol* 2000; *16:* 483–8.
33. Weisent J, Seaver W, Odoi A, Rohrbach B. Comparison of three time-series models for predicting campylobacteriosis risk. *Epidemiol Infect* 2010;*138:* 898–906.
34. Dominguez A, Munoz P, Cardenosa N, Martinez A, Cayla J. Time-series analysis of meningococcal disease in Catalonia. *Ann Epidemiol* 2007; *17:* 654–62.
35. Zhang X, Liu Y, Yang M, Zhang T, Young AA, Xiaosong L. Comparative study of four time series methods in forecasting typhoid fever incidence in China. *PLoS One* 2013; *8*(5) : e63116. doi:10.1371/journal.pone.006311.
36. Allard R. Use of time-series analysis in infectious disease surveillance. *Bull World Health Organ* 1998; *76*(4): 327–33.
37. Makridakis S, Wheelwright SC, Hyndman RJ. Forecasting: Methods and applications. III edn. New York: John Wiley & Sons Inc 1998; p. 370–8.
38. Riley S. Large-scale spatial-transmission models of infectious disease. *Science* 2007; *316* (5829): 1298–301.
39. Solomon PJ, Isham VS. Disease surveillance and data collection issues in epidemic modeling. *Stat Methods Med Res* 2000; *9*(3): 259–77.
40. Grassly NC, Fraser C. Mathematical models of infectious disease transmission. *Nature Reviews Microbiol* 2008; *6:* 477–87.
41. Hay SI, Were EC, Renshaw M, Noor AM, Ochola SA, Olusanmi I, *et al*. Forecasting, warning, and detection of malaria epidemics: A case study. *Lancet* 2003; *361*(9370): 1705–6.

Correspondence to: Dr Siraj Ahmed Khan, Scientist 'D', Regional Medical Research Centre-NER (ICMR), Post Box No. 105, Dibrugarh–786 001, Assam.
E-mail: sirajkhanicmr@gmail.com

Received: 17 Janaury 2014 *Accepted in revised form:* 3 April 2014