Applications of Self-Organising Map (SOM) for prioritisation of endemic zones of filariasis in Andhra Pradesh, India

Upadhayula Suryanaryana Murty*,
Mutheneni Srinivasa Rao, K. Sriram
and K. Madhusudhan Rao

Bioinformatics Group,
Biology Division,
Indian Institute of Chemical Technology (CSIR),
Tarnaka, Hyderabad 500 007,
Andhra Pradesh, India
Fax: 91-040-27193227
E-mail: murty_usn@yahoo.com
E-mail: msc_srinivas@yahoo.com
E-mail: palayam_k@yahoo.com
E-mail: raokadiri@yahoo.com
*Corresponding author

Abstract: Entomological and epidemiological data of Lymphatic Filariasis (LF) was collected from 120 villages of four districts of Andhra Pradesh, India. Self-Organising Maps (SOMs), data-mining techniques, was used to classify and prioritise the endemic zones of filariasis. The results show that, SOMs classified all the villages into three major clusters by considering the data of Microfilaria (MF) rate, infection, infectivity rate and Per Man Hour (PMH). By considering the patterns of cluster, appropriate decision can be drawn for each parameter that is responsible for disease transmission of filariasis. Hence, SOM will certainly be a suitable tool for management of filariasis. The detailed application of SOM is discussed in this paper.

Keywords: filariasis; data mining; SOMs; self-organising maps; prioritisation.

Reference to this paper should be made as follows: Murty, U.S., Rao, M.S., Sriram, K. and Rao, K.M. (2011) 'Applications of Self-Organising Map (SOM) for prioritisation of endemic zones of filariasis in Andhra Pradesh, India', *Int. J. Data Mining and Bioinformatics*, Vol. 5, No. 4, pp.417–427.

Biographical notes: Upadhayula Suryanaryana Murty is presently working as Scientist 'G'/ Director Grade Scientist, and Heading the Biology Division at Indian Institute of Chemical Technology (Council of Scientific and Industrial Research), Hyderabad. He has published over 80 research papers in peer reviewed journals in the areas of vectors and vector borne diseases, microbiology, bioinformatics and sericulture. His research interests are in medical entomology, bioinformatics applications, data mining and integrated pest management.

Mutheneni Srinivasa Rao is currently working as a scientist in Biology Division at Indian Institute of Chemical Technology (CSIR), Hyderabad since 2008. He received his Master's Degree in Biotechnology from Bharathidasan

University and PhD in Zoology from the Osmania University in 2008. His research interests are in epidemiological studies of vectors and vector borne diseases and disease modelling studies.

K. Sriram is presently working as a Technical officer, In Biology Division at Indian Institute of Chemical Technology (CSIR), Hyderabad. He received his Masters Degree in Zoology from M.P. BhoJ University, Bhopal. His research interests are in epidemiological studies of Lymphatic Filariasis.

K. Madhusudhan Rao is presently working as a Technical officer, In Biology Division at Indian Institute of Chemical Technology (CSIR), Hyderabad. He received his Masters Degree in Environmental Science from Kakatiya University and PhD in Zoology from Osmania University. His research interests are in epidemiological and socioeconomic studies of Lymphatic Filariasis.

1 Introduction

The Lymphatic Filariasis (LF) caused by *Wuchereria bancrofti* and spread by *Culex quinquefasciatus* mosquitoes is a major public health problem in Andhra Pradesh, India. LF is a socio-economic setback in almost 22 states/union territories of India, among which the state Andhra Pradesh contributes about 10% of filariasis load. Since many decades, it experienced the difficulty to tackle LF, which is mainly due to improper disease surveillance and ineffective control management of mosquitoes and *Wuchereria bancrofti* (Ramaiah et al., 2000; Snehalatha et al., 2003). There is an urgent need for the implementation of appropriate tools/models to identify the endemic regions for monitoring and evaluation of effective control programme.

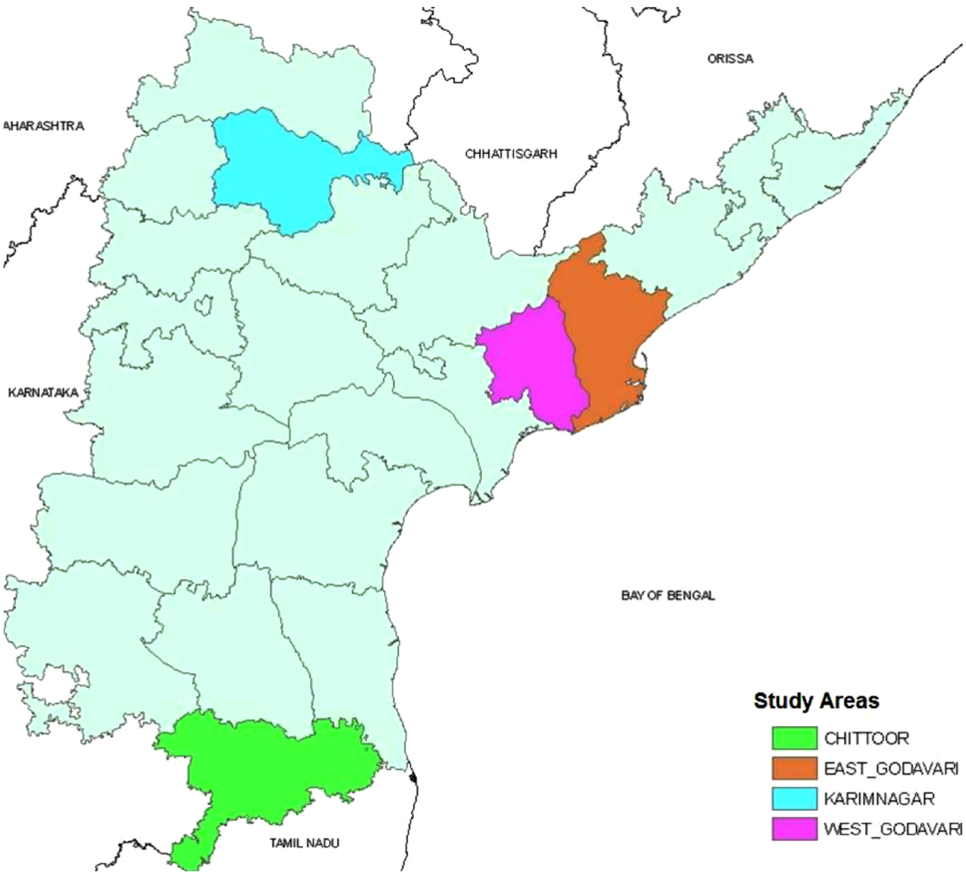
It is evident that judgement on frequency of filarial cases in future is very critical. In such situations, development of a relevant model for infectious diseases is necessary for understanding the filariasis status in the targeted region and seems to be the proper way to suppress the disease burden. To stress the point, application of mathematical/statistical models and computational modelling of infectious diseases can provide key inputs like population dynamics of parasite, prediction of epidemiological trend and aiding decision about control strategies. Applications on computational modelling in disease management of Cardiovascular Risk (Pfaff et al., 2004), prediction of malaria in Kenya (Githeko and Ndegwa, 2001), uses of Bayesian Network for control of malaria (Cancre et al., 2000), data-mining applications for the control of nosocomial infections and antimicrobial resistance studies (Brossette et al., 2000) have been well described by various researchers. SOM, a data-mining tool, has been widely applied in various engineering applications of pattern recognition, full-text and image analysis, vector quantisation, regression, financial data analysis and fault diagnosis (Costa et al., 2001; Jin et al., 2001; Kohonen et al., 2000). The output of SOM gives clustering patterns, which have been applied to a wide range of classification problems in public health and other related fields. This cluster analysis provides quantitative framework for cross study and interregional comparison of related factors. Earlier, we have also reported the application of SOM to prioritise the malaria endemic zones in Manipur state of India (Murty et al., 2008). Similar type of approach has also been applied in this present investigation to prioritise the filariasis endemic zones in four districts of Andhra Pradesh.

2 Materials and methods

2.1 Study areas

The state of Andhra Pradesh, India, is administratively consisting of 23 districts, out of which 13 districts are reported endemic for filariasis. Out of these 13 districts, 4 districts (Karimnagar, Chittoor, East and West Godavari) were selected covering 120 villages (Figure 1). Thirty villages from each district Karimnagar and Chittoor, 45 villages from East Godavari and 15 villages from West Godavari district were selected by stratified randomly sampling with the support of the district health officials during 2004–2007. From all the villages, data was collected for epidemiological and entomological parameters following the standard protocols on filariasis.

Figure 1 Map showing study areas (Districts) in Andhra Pradesh (see online version for colours)



Source: Maps of India

2.2 Data source and variables for 'SOM'

Data on Microfilaria (MF) rate, PMH, infection and infectivity rate were made use for cluster analysis using SOM. Mosquito collection was carried out from all 120 villages to quantify the density of mosquitoes (PMH). The collected female *Culex quinquefasciatus* mosquitoes were dissected individually to quantify the infection and infectivity rates. MF rate from the individuals was assessed by the examination of blood smears collected from consenting individuals by stratified random sampling method. The parameters were calculated by using the following formulas:

$$\text{Per Man Hour} = \frac{\text{No. of female mosquitoes collected}}{\text{The time spent}}$$

$$\text{Microfilaria rate (\%)} = \frac{\text{Positive blood smears}}{\text{Total blood smears examined}} \times 100$$

$$\text{Infection rate (\%)} = \frac{\text{No. positive for } L_1, L_2, L_3}{\text{No. of mosquitoes dissected}} \times 100$$

$$\text{Infectivity rate (\%)} = \frac{\text{No. of mosquitoes +ve for } L_3 \text{ stage}}{\text{No. of mosquitoes dissected}} \times 100.$$

2.3 Self-Organising Maps (SOMs)

SOM is a data-mining tool for the visualisation of high-dimensional data in a two-dimensional manner, and the creation of abstractions like in many clustering techniques. It has the characteristic nature to analyse the inputs with similar cells and also with neighbourhood cells that contain similar types of inputs. This property together with the easy visualisation makes the SOM a useful tool for visualisation and clustering of large data sets. Output of SOM result helps to prioritise the parameters in a more appropriate way so that effective decision can be taken on the problem. The detail algorithm of SOM is mentioned in our earlier study (Murty et al., 2008).

An SOM layer of 3×3 neurons can represent approximately 1/10 of 100 data points in a data set resulting in 9 clusters that contain similar data points (from a physical viewpoint) whereas their class labels might be dissimilar (low specificity or sensitivity) or similar (high specificity or sensitivity) according to a given benchmark and frequency threshold. Hence, each neuron represents a smaller set of physical similar data points. Thus, the number of class members and the class frequency of this cluster (neuron) change as well as sensitivity and specificity values. Finally, these class frequency changes can alter the results significantly.

2.4 Standardisation of data

The summarised data has been standardised between 0 and 1 and all the parameters (i.e., MF rate, PMH, Infection and Infectivity rates) have been given equal importance while clustering. The neuron weightage has been adjusted by the learning rate; the learning rates and distance threshold values for the SOM are generally default values (Markey et al., 2003). Unsupervised learning was done using the data learning constant of

0.01–0.1 with 5000 iterations that yielded clusters based on the neighbourhood distance. The mean values comparison, 95% confidence intervals and significance differences (*p*-values) among infection, infectivity rates, PMH and MF rates were analysed and represented in Tables 1–4.

Table 1 Mean and other statistical values for infection rate – by districtwise

Name of the district	*Sample size*	*Infection rate*		
		Mean	*95% confidence interval*	*p-value*
Karimnagar	30	11.082	7.694–14.470	0.000
Chittoor	30	4.062	1.378–6.746	0.004
East Godavari	45	13.851	11.975–15.727	0.000
West Godavari	15	14.680	12.600–16.760	0.000

Table 2 Mean and other statistical data for infectivity rate – by districtwise

Name of the district	*Sample size*	*Infectivity rate*		
		Mean	*95% confidence interval*	*p-value*
Karimnagar	30	1.975	0.781–3.169	0.002
Chittoor	30	0.890	–0.036–1.816	0.059
East Godavari	45	2.894	2.340–3.449	0.000
West Godavari	15	2.137	1.563–2.711	0.000

Table 3 Mean and other statistical data for MF rate – by districtwise

Name of the district	*Sample size*	*MF rate*		
		Mean	*95% confidence interval*	*p-value*
Karimnagar	30	2.067	1.166–2.968	0.000
Chittoor	30	0.883	0.290–1.477	0.005
East Godavari	45	10.982	9.085–12.879	0.000
West Godavari	15	22.713	18.040–27.386	0.000

Table 4 Mean and other statistical data for PMH by districtwise

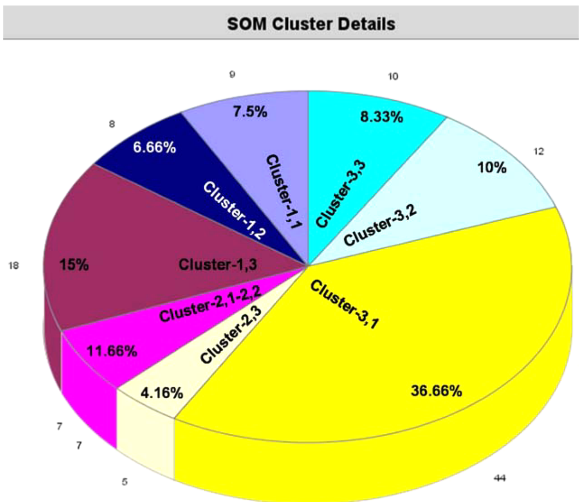
Name of the district	*Sample size*	*PMH*		
		Mean	*95% confidence interval*	*p-value*
Karimnagar	30	34.400	30.298–38.502	0.000
Chittoor	30	30.845	28.461–33.228	0.000
East Godavari	45	39.436	36.271–42.600	0.000
West Godavari	15	35.620	32.205–39.035	0.000

3 Results and discussion

SOM classified the data into hyper (red), hypo (maroon) and low (black) as endemic or non-endemic regions of filariasis at learning rate from 0.1 to 0.01 for each parameter.

The study area of 120 villages was categorised into 3 major clusters (9 sub-clusters) out of which Cluster 1 group comprising (1.1, 1.2 and 1.3) represented by 35 villages (Figure 2) are found to be highly endemic for filariasis and Cluster 2 (2.1, 2.2 and 2.3) comprising 19 villages (Figure 2) are observed to be moderately endemic filarial zones, whereas Cluster 3 classified as (3.1, 3.2 and 3.3) with inputs from 66 villages (Figure 2) have been classified as non-endemic or least prone zones for filariasis. By observing the various types of clusters, derived by SOM software, it is clearly understood that the intensity of filaria largely varied among the villages where the study was undertaken. Among the 4 districts, study villages from East Godavari district were found to be highly endemic for filariasis. The parameters influencing the onset of disease like PMH of *Culex quinquefasciatus*, Microfilaria rate, infection and infective rates also varied among the villages of Clusters (1, 2 and 3). It is noticed that, among these four parameters, each parameter has significant contribution for the filarial cases in a particular village duly supported by the presence of other three parameters responsible for the disease.

Figure 2 Pie chart of SOM clusters (3 × 3) (see online version for colours)



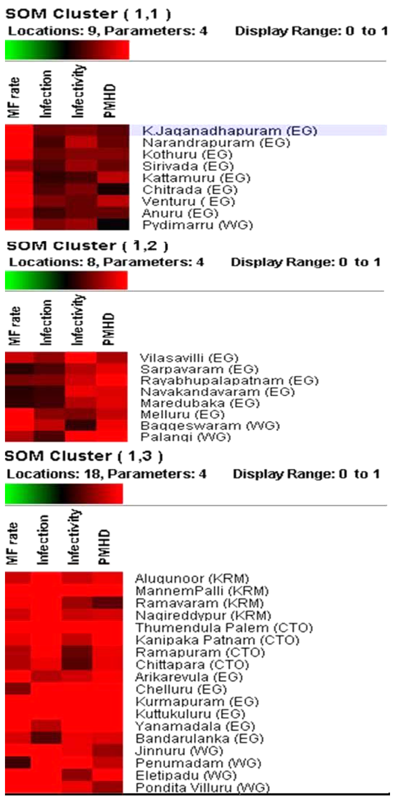
3.1 Cluster 1 (1.1, 1.2 and 1.3)

Cluster 1 has categorised 35 villages among 120 surveyed villages, into 3 sub-clusters (Clusters 1.1, 1.2 and 1.3). Nine villages, which are represented in Cluster 1.1 with 7.5% Microfilaria rate, were found to be alarming, whereas the other parameters like infection, infectivity and PMH were found to be very minimal when compared with the MF rate (Figure 3). Hence, it is assumed that occurrence of filarial disease is very common in these villages represented by Cluster 1.1 and the population has been already infected with microfilaria. The lower density of infection, infectivity and PMH noticed in these villages may be due to the successful vector control operations carried out by the authorities in these villages. However, it indicates that very less attention might have been paid to eradicate the MF rate from the population. As the individuals in these villages have already acquired the MF, there is every possibility of transmission of disease in future as there will be exponential increase in infected mosquito population.

Hence, in these affected villages, a sound decision on MF control operations like Mass Drug Administration (MDA) is more appropriate rather than mosquito control operations. The manpower (health officials) and material (medicines for control measures) that is supposed to be imparted for mosquito control can be relaxed in these villages and the resources can be effectively imposed on drug administration to check the microfilaria rate so that transmission of the disease can be arrested.

Similarly, in Cluster 1.2, 8 villages (66%) are found to be endemic with high infectivity rate and mosquito density (PMHD). Although the density of MF and infection rate are found to be very low in these villages, less filarial transmission is noticed in this cluster of villages, which may be due to the occurrence of very minimal vector human contact. Among the eight villages of this cluster, high MF and infectivity rates were recorded in Melluru and Baggeswaram, which suggest that MF transmission in these two villages has already occurred. Hence, in these villages, drug distribution is to be given precedence over the other control operations to kill the parasites. In the other villages due to the high incidence of infectivity rates and PMHD, precaution should be prescribed to avoid the mosquito human contact for further transmission of the MF parasite. If vector control operations are not employed effectively in these villages, there is every possibility that the infective mosquitoes will spread the MF transmission in due course of time. Hence, it is imperative that a decision should be taken for implementation on priority basis for operations like anti-larval and anti-adulticide to reduce the vector density (PMHD) in this cluster of villages.

Figure 3 3 × 3 SOM Cluster (1.1, 1.2 and 1.3) (see online version for colours)

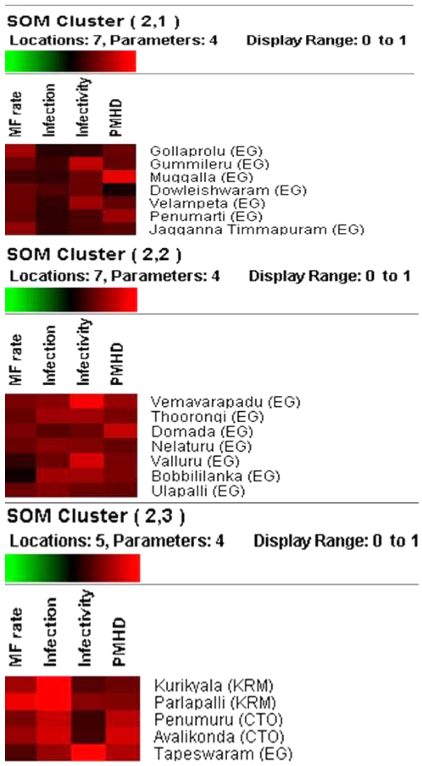


In Cluster 1.3, 18 villages are classified (15%) and it is found that all the villages are highly endemic for filariasis. The results indicate that no control measures might have been implemented in these villages due to which high PMHD, infection, infectivity and MF rates were noticed in all the villages. So, in this case, control operations should be focused intensively towards MDA as well as vector control (larval and adult) to suppress the MF rate and vector density simultaneously.

3.2 Cluster 2 (2.1, 2.2 and 2.3)

Cluster 2 is grouped as (2.1, 2.2 and 2.3) into 19 villages (11.6%, 11.6% and 4.16%) and all the villages coming under these clusters are found to be less endemic for filariasis (Figure 4). These regions have relatively low MF rates, and vector abundance also varied from moderate to low levels. The infection and infectivity rates are also found to be moderately high in Clusters 2.2 and 2.3. Hence, in these villages, it is suggested that health officials should focus more on controlling the mosquito density to achieve effective results in arresting the transmission of the filarial parasite.

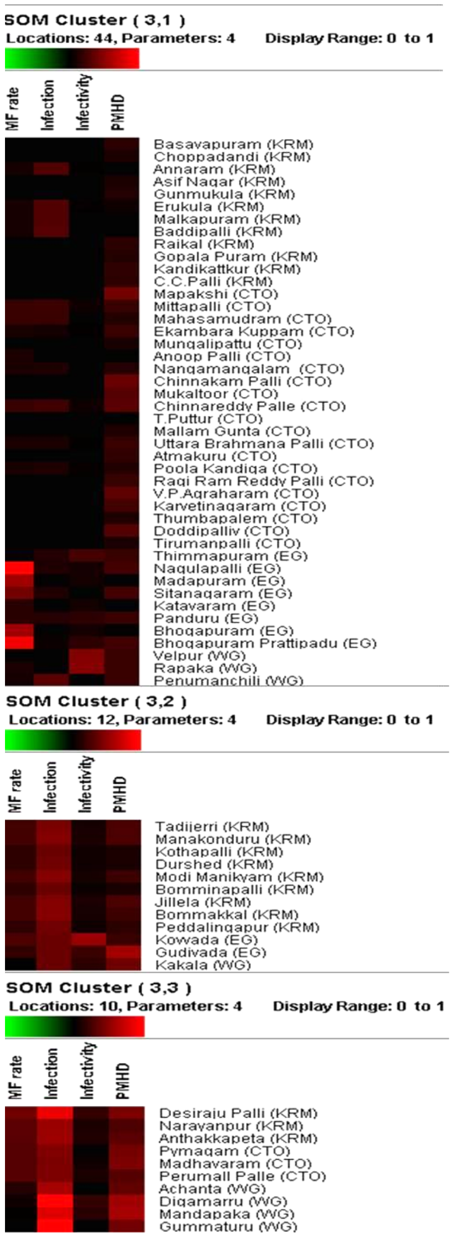
Figure 4 3 × 3 SOM Cluster (2.1, 2.2 and 2.3) (see online version for colours)



3.3 Cluster 3 (3.1, 3.2 and 3.3)

Cluster 3 is classified as (3.1, 3.2 and 3.3) with inputs collected from 66 villages in the percentage of representation of the following order (36.66%, 10% and 8.33%) (Figure 5). It is found that the MF rates in these villages are exceptionally low and alarming situation is observed only in Nagulapalli and Bhogapuram Prattipadu of East Godavari district. It is suggested that in these villages the MDA should be implemented on priority basis. Similarly, the infection rates recorded were moderately high in the villages represented in Clusters 3.2 and 3.3. It is found suitable that in this case immediate emphasis should be given on the vector control operations to achieve good results.

Figure 5 3 × 3 SOM Cluster (3.1, 3.2 and 3.3) (see online version for colours)



4 Conclusion

Lymphatic Filariasis is one of the dreaded diseases plaguing the entire South East Asia, with large contribution of diseased patients from India. Andhra Pradesh is one of the states that is bearing this burning problem and especially more prevalent in the coastal regions where the breeding places for the mosquitoes are aplenty, the suitable ecoclimatic conditions favouring the *Culex* mosquitoes and the nematode parasite, the presence of large number of persons harbouring the parasites for transmission to healthy individuals through vectors.

Though systemic control operations are being carried out regularly, still there is a huge communication gap between the grass root workers and the administrators, creating a very critical situation in understanding the actual dynamics of the disease transmission. At this juncture, it is very important to have a proper understanding of the distribution of disease and targeting the filarial infection in the endemic regions. The explosion of information technology has opened many avenues and data mining is one of them. By making effective use of "Self Organising Maps" (SOMs), prioritisation of the filarial endemic regions can be attained as per the severity of the parameters governing the disease in the respective villages. Once prioritisation is done, the control operations can be launched effectively, with the knowledge gained by SOM all the villages can be mapped in a more precise way. This will help to understand the nature of disease dynamics in the endemic regions, so that public health authorities can concentrate on a specific parameter to control the disease more effectively.

Acknowledgements

The authors are grateful to Director, Indian Institute of Chemical Technology, for his encouragement and support. The authors are thankful to Ministry of Communication and Information Technology, Govt. of India, New Delhi, for sponsoring and supporting the Project.

References

Brossette, S.E., Sprague, A.P., Jones, W.T. and Moser, S.A. (2000) 'A data mining system for infection control surveillance', *Methods of Information in Medicine*, Vol. 39, pp.303–310.

Cancre, N., Tall, A., Rogier, C., Faye, J. and Sarr, O. (2000) 'Bayesian analysis of an epidemiologic model of *Plasmodium falciparum* malaria infection in Ndiop, Senegal', *American Journal of Epidemiology*, Vol. 152, pp.760–770.

Costa, J.A.F. and Andrade Netto, M.L. (2001) 'A new tree-structured self-organizing map for data analysis', *Proceedings of International Joint Conference on Neural Networks*, Vol. 3, pp.1931–1936.

Githeko, A.K. and Ndegwa, W. (2001) 'Predicting malaria epidemics in the Kenyan highlands using climate data: a tool for decision makers', *Global Change and Human Health*, Vol. 2, pp.54–63.

Jin, H., Leung, K.S. and Wong, M.L. (2001) 'An integrated self-organizing map for the travelling salesman problem', in Mastorakis, N. (Ed.): *Advances in Neural Networks and Applications*, World Scientific and Engineering Society Press, pp.235–240.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V. and Saarela, A. (2000) 'Organization of a massive document collection', *IEEE Transactions on Neural Networks,* Vol. 11, pp.574–585.

Markey, M.K., Joseph, Y.L., Tourassi, G.D. and Floyd, C.E. (2003) 'Self-organizing map for cluster analysis of a breast cancer database', *Artificial Intelligence in Medicine*, Vol. 27, pp.113–127.

Murty, U.S.N., Rao, M.S. and Misra, S. (2008) 'Prioritization of malaria endemic zones using self-organizing maps in the Manipur state of India', *Informatics for Health and Social Care*, Vol. 33, pp.170–178.

Pfaff, M., Weller, K., Woetzel, D., Guthke, R., Schroeder, K., Stein, G., Pohlmeier, R. and Vienken, J. (2004) 'Prediction of cardiovascular risk in hemodialysis patients by data mining', *Methods of Information in Medicine*, Vol. 43, pp.106–113.

Ramaiah, K.D., Das, P.K., Michael, E. and Guyatt, H. (2000) 'The economic burden of lymphatic filariasis in India', *Parasitol Today*, Vol. 16, pp.251–253.

Snehalatha, K.S., Ramaiah, K.D., Vijay Kumar, K.N. and Das, P.K. (2003) 'The mosquito problem and type and costs of personal protection measures used in rural and urban communities in Pondicherry region, South India', *Acta Tropica*, Vol. 88, pp.3–9.