Prioritization of Malaria endemic zones in Arunachal Pradesh: A novel application of self organizing maps (SOM)
U Muty, N Arora

Citation

Abstract
Malaria continues to pose a serious threat to public health in North-Eastern states of India. Arunachal Pradesh is highly endemic for Malaria predominately with Plasmodium falciparum infections. Despite continuous efforts by government, a desirable level of control has not been achieved. The present study describes the application of self organizing maps (Kohonen maps), a data mining tool for prioritization of malaria endemic zones in this region. 60 PHCs (Public Health Centers) were randomly selected from Arunachal Pradesh and 6 maliariometric parameters via Annual Blood Examination rate (ABER), Annual Parasite Incidence (API), Slide Positivity Rate (SPR), Annual Falciparum Incidence (AFI) and Slide Falciparum Rate (SFR) were considered which reflected the intensity of malaria transmission in this region. Self Organizing Maps yielded 9 clusters based on neighborhood distance, which reflects about zones based on status of intensity of malaria epidemiology. Such maps would make it possible to target control measures at high-risk areas and greatly increase the cost efficiency of malaria control programmes.

INTRODUCTION
Malaria, the third leading cause of death attributable to an infectious disease worldwide, has plagued mankind for countless generations. Malaria remains a public health problem in 90 countries in the world (1) and causes more than 300 million acute illnesses and at least one million deaths annually (2). The annual malaria burden in India estimates to nearly 2 to 2.5 million cases. North-Eastern region of India is in the Indo-Chinese hill zone of Macdonald's classification of stable malaria (3) and contributes nearly 9% of total malaria cases in India (4). In this region, efficient malaria transmission is maintained during most months of the year and slashes potential economic growth and thus is a major impediment to the overall development and progress of these areas.

Despite of several anti-malaria programmes being implemented under National Vector Borne Diseases Control Programme, this region has seen little tangible progress in alleviating the burden of malaria (5, 6). Apparently, there are definite inadequacies that continue to dampen the spirit of public health specialists even during the halcyon days of malaria eradication.

On closer scrutiny, it was evident that, there being financial and technical constraints common to all states of India, operational difficulties are hampering the effective malaria control in the North-Eastern region (7). These very areas remain inaccessible owing to floods and poor road communication. The major reasons of perennial and persistent malaria transmission are predominance of Plasmodium falciparum (8, 9), difficult terrain (10), congenial eco-climatic conditions (11), lack of proper implementation of control operations and ineffective communication between health researchers and policy makers. The problem of drug resistance (11, 12, 13), exophilic and exophagic vector behaviour and high efficiency of vectors (11, 14, 15) further aggravate the situation. Due to these various factors encountered in the North-Eastern region, malaria continues to present health services with an immensely difficult and complex challenge. The highly focal nature of malaria requires targeting of interventions to specific regions and malaria control interventions must be preceded by the identification and prioritization of the most vulnerable. Hence, there is an imperative need for exploitation of advanced Information Technology tools which can prioritize the endemic zones to liberate the region from manacles of this pandemic. Information technology has been successfully
exploited in different spheres of control of vector-borne diseases. Computer applications in database management and data mining have paved way for control of malaria (16) and Filariasis (17, 18) and have proved to be a valuable tool in decision making in vector identification (19). Computer simulation models of vector borne diseases have been used to develop early warning system for the epidemics (19). Various expert systems have also been developed employing Artificial Intelligence (AI) for insect pest management and for forecasting of outbreak of vectors. SOMs (originally developed by Teuvo Kohonen) are special architecture of neural networks or special types of Artificial Neural Networks (ANN) that cluster the high dimensional data vectors according to a similarity measure. SOMs have successfully been applied for classification of DNA sequences based on codon usage (20, 21), nucleotide frequencies (22), virtual potentials (23), protein sequences analysis (24, 25) and clustering of microarray data (26). SOM have been successfully exploited in various epidemiological studies (27, 28, 29) In accordance with their important role in easy visualization of complex epidemiological data, Self Organizing Maps (SOM) are customized and used in the present study to prioritize the malaria endemic zones for disease management in Arunachal Pradesh.

MATERIALS AND METHODS

Study site: Arunachal Pradesh is the largest state area-wise situated in the North-East region of India, sharing a long international border with Bhutan, China and Myanmar. This state is situated between latitude 26 30' N and 29 30' N and longitude 91 30' E and 97 30' E. The climate of the state is dominated by the Himalayan system and the altitudinal variations. The climate is highly hot and humid at the lower altitudes and in the valleys covered by swampy dense forest particularly in the eastern section, while it becomes exceedingly cold in the higher altitudes. Average temperature during the winter months ranges from 150 C to 210C and 220C to 300C during monsoon. Forested terrain and perennial streams are congenial for rapid multiplication and longevity of malaria vectors. Population of state is estimated to be 1091117 according to 2001 census. The state has a major population of 20 scheduled tribes and numerous sub-tribes. Agriculture is the primary driver of the economy. Nearly 80% of the population is engaged in agriculture. The traditional method of agriculture is Jhumming, a kind of shifting cultivation. The main crops are rice, maize, millet, wheat and mustard.

METHODS

Data Collection: Raw data was collected from the Directorate of Health, Govt. of Arunachal Pradesh, which consists of Epidemiological aspects of Malaria cases encountered in 60 randomly selected Public Health Centers belonging to 12 districts of Arunachal Pradesh in 2005. Raw data pertaining to malaria incidence was collected and standard malarialmetric parameters (ABER, API, SPR, SFR, AFI) were calculated based on this data to be used in this study. (Table1)
Figure 1

Table 1: Malarometric indices of Public Health centers in Arunachal Pradesh

<table>
<thead>
<tr>
<th>Name of PHC</th>
<th>ABER</th>
<th>API</th>
<th>SPI</th>
<th>API</th>
<th>SPI</th>
<th>API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Papaghat</td>
<td>98.73</td>
<td>41.09</td>
<td>66.31</td>
<td>42.15</td>
<td>31.06</td>
<td>45.84</td>
</tr>
<tr>
<td>Kalingo</td>
<td>91.69</td>
<td>31.06</td>
<td>62.64</td>
<td>32.13</td>
<td>21.71</td>
<td>34.73</td>
</tr>
<tr>
<td>Roing</td>
<td>90.84</td>
<td>39.62</td>
<td>59.72</td>
<td>31.46</td>
<td>20.80</td>
<td>33.77</td>
</tr>
<tr>
<td>Tawang</td>
<td>90.30</td>
<td>39.83</td>
<td>59.21</td>
<td>31.13</td>
<td>20.57</td>
<td>33.39</td>
</tr>
<tr>
<td>Bonggo</td>
<td>89.37</td>
<td>38.92</td>
<td>57.91</td>
<td>30.62</td>
<td>19.84</td>
<td>32.25</td>
</tr>
<tr>
<td>Nagaland</td>
<td>88.61</td>
<td>38.63</td>
<td>56.80</td>
<td>30.15</td>
<td>19.23</td>
<td>31.71</td>
</tr>
<tr>
<td>Basar</td>
<td>87.95</td>
<td>38.63</td>
<td>56.31</td>
<td>30.15</td>
<td>18.39</td>
<td>31.37</td>
</tr>
<tr>
<td>Maningyok</td>
<td>87.21</td>
<td>38.42</td>
<td>55.83</td>
<td>29.67</td>
<td>17.81</td>
<td>30.95</td>
</tr>
<tr>
<td>Tawangnak</td>
<td>86.48</td>
<td>38.23</td>
<td>55.37</td>
<td>29.19</td>
<td>17.29</td>
<td>30.54</td>
</tr>
<tr>
<td>Namsai</td>
<td>85.75</td>
<td>38.04</td>
<td>54.92</td>
<td>28.73</td>
<td>16.80</td>
<td>29.12</td>
</tr>
<tr>
<td>Kekching</td>
<td>84.94</td>
<td>37.86</td>
<td>54.49</td>
<td>28.27</td>
<td>16.31</td>
<td>28.70</td>
</tr>
<tr>
<td>Namsukung</td>
<td>84.13</td>
<td>37.67</td>
<td>54.08</td>
<td>27.84</td>
<td>15.83</td>
<td>28.29</td>
</tr>
<tr>
<td>Bojo</td>
<td>83.32</td>
<td>37.49</td>
<td>53.68</td>
<td>27.41</td>
<td>15.36</td>
<td>27.88</td>
</tr>
<tr>
<td>Sajeev</td>
<td>82.51</td>
<td>37.31</td>
<td>53.29</td>
<td>26.99</td>
<td>14.90</td>
<td>27.48</td>
</tr>
<tr>
<td>Roing</td>
<td>81.70</td>
<td>37.14</td>
<td>52.91</td>
<td>26.59</td>
<td>14.44</td>
<td>27.09</td>
</tr>
<tr>
<td>Basu</td>
<td>80.89</td>
<td>36.98</td>
<td>52.53</td>
<td>26.20</td>
<td>13.99</td>
<td>26.70</td>
</tr>
<tr>
<td>Ambukuk</td>
<td>80.08</td>
<td>36.81</td>
<td>52.16</td>
<td>25.82</td>
<td>13.54</td>
<td>26.31</td>
</tr>
<tr>
<td>Namsukung</td>
<td>79.27</td>
<td>36.65</td>
<td>51.79</td>
<td>25.46</td>
<td>13.10</td>
<td>25.93</td>
</tr>
</tbody>
</table>

Note: ABER, API, SPI, SPI, API, PPI are indices used to prioritize malaria endemic zones.
DATA ANALYSIS: DATA MINING – SELF ORGANIZING MAPS

In SOM, neurons compete with each other to earn the right of representing the input data \((30, 31)\). As a result, data in the multidimensional attribute space can be abstracted to a much smaller number of latent dimensions organized on a basis of a predefined geometry in a space of lower dimensionality, usually a regular two-dimensional array of neurons. Via this way the structures embedded in the input data can be revealed which is placed in the input space and is spanned over the inputs distribution. Using a SOM network, it is possible to obtain a map of input space where closeness between units or clusters in the map represents closeness of the input data. Processing units in the SOM lattice is associated with weights of the same dimension of the input data. Using the weights of each processing unit as a set of coordinates the lattice can be positioned in the input space. During the learning stage the weights of the units change their position and “move” towards the input points. This “movement” becomes slower and at the end of the learning stage, the network is “frozen” in the input space. After the learning stage the inputs can be associated to the nearest network unit. When the map is visualized, the inputs can be associated to each cell on the map. One or more cell that clearly contains similar objects can be considered as a cluster on the map. These clusters are generated during the learning phase without any other information. It is not necessary to supply to the network cluster prototypes or examples. SOMs cluster the data in a manner similar to cluster analysis, but have an additional benefit of ordering the clusters and enabling the visualization of large numbers of clusters. These clusters are arranged in a low-dimensional topology-usually a grid structure that preserves the neighborhood relations in the high dimensional data \((32, 33)\). This technique is particularly useful for the analysis of large datasets where similarity matching plays a very important role \((34)\). The characteristic that distinguishes the SOM net from the other classification algorithms is that not only similar inputs are associated to the same cell but also neighborhood cells contain similar documents. This property together with the easy visualization makes the SOM map a useful tool for visualization and clustering of large data sets.

Parameters identified for SOM: Parameters like Annual Blood Examination rate (ABER), Annual Parasite Incidence (MPI), Slide Positivity rate (SPR), Slide Falciparum rate (SFR) and Annual Falciparum Rate (AFR) and Plasmodium falcipraum % were considered for this study. All the factors were given equal importance.

Data Normalization: Summarized data is normalized linearly such that minimum value in each category is 0 and the maximum 1. This is done to ensure that all the parameters are given equal importance when clustering is done.

RESULTS

Normalized data is clustered using SOM yielded 9 clusters on a 3x3 (shown in figure1). Unsupervised learning was done on the fly using the data using a learning constant of 0.01 and for 10,000 iterations following which the data got clustered among clusters based on the neighborhood distance.

Figure 2
Normalization Formulae = \( \frac{\text{Original data value} - \text{Minimum Data value}}{\text{Maximum data value} - \text{Minimum Data value}} \)

Figure 3

Figure 1: SOM Clusters showing different endemicity levels of Public Health centers.

DISCUSSION

Cluster (1, 1), Cluster (1, 2) and Cluster (1, 3): These PHCs shows very low ABER and thus cuts a poor picture where because of the lack of appropriate surveillance, there is a possibility of underestimating the disease burden. There is an immediate need to investigate the causes hindering the proper survey and strengthening the HealthCare infrastructure. In spite of moderate SPR observed, some regions like Jairampur, Bordumsa and Bhalukpong are
showing very high Pf% and thus requires immediate attention to curb this deadly parasite from its transmission to other regions.

**Figure 4**

Cluster (2, 1): Though the PHCs under this cluster shows low to moderate incidence of malaria yet a very high Pf% immediates the need of focusing on the drug administration on priority.

**Figure 6**

Cluster (2, 2): Only one PHC Khimiyong fell in this cluster and indicates moderate incidence of malaria. This region requires both drug administration and vector control measures at relatively less priority.
Prioritization of Malaria endemic zones in Arunachal Pradesh: A novel application of self organizing maps (SOM)

Figure 7

Cluster (2, 3): Low to moderate AP and SPR was observed in the 10 PHCs under this cluster and almost nil falciparum infections indicates that drug administration and vector control operations can be done with less priority.

Figure 8

Cluster (3, 1): The PHCs in this cluster shows an alarming rate of high malariometric indices. Seppa shows very high level of all malariometric parameters and reflects high endemicity. The PHCs clustered in this cluster are showing predominance of P. falciparum and thus drugs particularly targeting drug resistant falciparum malaria should be administered in this region. Inspite of efforts and various control operations undertaken by NMEP, malaria is still deeply entrenched in these regions. These regions warrant further investigation and more focused efforts on active surveillance and require both drug administration and vector control measures at relatively high priority and need a radical overhaul in the way it tackles the disease.

Figure 9

Cluster (3, 2): 5 PHCs clustered in this region shows moderate level of all malariometric indices and clearly reflects less falciparum trend as indicated by SFR, AFI and PF%. Only Sille shows very high API, and thus needs malaria control interventions at priority.

Figure 10

Cluster (3, 3): This cluster included the regions where due to efficient efforts of government machinery, ABER is considerably high which will help in getting a clear picture of status of malaria in the regions and will truly be a
reflection of surveillance. These regions demonstrate a very high annual Parasite incidence (API) and moderate to high SPR but very low or almost negligible falciparum incidence is reflected by SFR and AFI.3 PHCS namely Telam, Rani and Nari are demonstrating very high API This shows that despite highly satisfactory surveillance leading to high ABER malaria was still refractory to intervention measures and hence, there is a need to excogitate our strategy for control again.

CONCLUSIONS
The application of Data mining and artificial intelligence in Epidemiology is still in its infancy. In spite of numerous evidences of incorporation of artificial intelligence as an aid in data analysis of various epidemiological studies, medical entomologists are still unable to tap its potential in vector control except for data acquisition and storage. Information Technology in vector control operations has been extended to construction of Databases on different aspects of vector borne diseases and various forecasting systems based on computer simulation models \( \phi \). Application of Artificial Intelligence in combating vector borne diseases can give a completely new dimension to existing control programs. Artificial neural Networks such as Kohonen Maps have a natural propensity to learn—they learn how to solve problems from data as opposed to solving problems based on explicit problem specification \( \phi \). Self Organizing maps (SOM) are deemed as being highly effective as a sophisticated visualization tool for visualizing high dimensional complex data with inherent relationships between the various features comprising the data. These have been successfully exploited in Medical and Health Informatics in fields as varied as Medical image processing \( \psi \), disease diagnosis \( \phi \), gene prediction \( \phi \), gene sequence analysis \( \phi \), structural recognition of protein families \( \phi \) and drug designing \( \phi \) and drug utilization \( \phi \). In recent past, SOMs have been employed for data exploration in major public health diseases like Diabetes \( \phi \), Glaucoma \( \phi \). In this paper, we have shown the use of Self Organizing Maps as valuable tool in prioritization of malaria endemic zones which will assist in decision making on the location and deployment of health care services and prioritization of intervention strategies. In areas like Arunachal Pradesh which suffer from perennial malaria transmission, and where difficult terrain and geographical features are big hurdles in carrying out effective and timely vector control operations, SOM will be very effective in bridging the gap between policy makers and Health workers. Recognizing consistent foci of cases would permit control efforts to be directed at specific geographic areas, reducing costs and increasing effectiveness. In a country like ours where resources are scarce, reliable methods for the stratification of zones on basis of the prevalence or transmission intensity of malaria are urgently required. Such clustering and data visualization tools are essential for assessing the austerity of the problem, and hence the resources needed to emulate malaria. This approach will serve as yardstick for assessing the progress of control and indicate which geographic areas should be prioritized, so that large amount of man power and resources can be saved. Because of underlying simplicity in data visualization, SOM will prove to be a powerful weapon in arsenal in fight against this dreaded disease. This strategy will play a crucial role in bridging research and control and it is quite likely that besides reducing the malaria burden, the entire public health system will benefit from such a strategy if adopted and extrapolated to other regions across the world for other vector borne diseases.

ACKNOWLEDGEMENTS
Authors are grateful to the Director, IICT, Hyderabad for his continuous support and encouragement. Neelima Arora thanks CSIR for Senior Research Fellowship.

References
10. Yadava RL, Sharma RS. Malaria problem and its control
Prioritization of Malaria endemic zones in Arunachal Pradesh: A novel application of self organizing maps (SOM)

Author Information

U.S.N. Muty, Ph.D.
Scientist E1, Deputy Director, Head, Biology Division, Indian Institute of Chemical Technology

Neelima Arora, M.Sc.
Senior Research Fellow, Biology Division, Indian Institute of Chemical technology