Modeling the influence of socio-economic factors on HIV prevalence in India using Artificial Neural Network on spatial database

Abstract: The proliferation of spatial information during last several years has attracted the policy makers to extract useful patterns from large spatial datasets[1][2][3]. The heterogeneous geographical distribution of HIV/AIDS epidemiology and varied spatial socio-economic factors led us to work out an intelligent model which is able to correlate the trend of spread of this disease in India. In this paper we empirically study the role of socio economic factors like migration rate[MGRT], ratio of female to male literacy[LTFM], average distance traveled by migrant/bridge population[AIMD], human development index [HDI], gender development index [GDI], in populating the disease in India. An Artificial Neural Network based model has been developed which has correlated these spatial and non-spatial factors with the various spread pattern of the disease. The result of the model reveals an interesting pattern which in agreement with the report published by the government on the basis of the physical survey of various geographical locations.

Key Words: Spatial Data Base, Spatial Data Mining, ANN-BPN Model, Oracle10g.

1. Introduction
Spatial data mining has become a new and powerful tool for efficient and complex analysis of very large geospatial database [4][5][6]. It puts emphasis on extraction of interesting and implicit knowledge such as the spatial pattern or other significant mode not explicitly stored in the spatial database[7][8]. The geographical attributes involved in spatial database is an important aspect for many applications[4][9][1]. The HIV sentinel surveillance obtains HIV prevalence data from antenatal clinics (ANC) and sexually transmitted disease (STD) clinics as well as from high-risk groups. The information contained in the sentinel surveillance database reveals that India has a heterogeneous HIV epidemic. The main objective of this research is to develop an intelligent model which is able to take into account the important socio economic factors and forecast the prevalence, growth or declining trend of the epidemic like HIV/AIDS at various geographical locations[4][5]. The knowledge given by the model can be used to perform spatial prediction that could help the policy makers to plan and monitor the impact of HIV prevention and care intervention program.

The conventional analysis techniques have been based on traditional statistics and multidimensional data analysis. The
traditional analysis is performed by diverse method like basic statistics (average, variance, histogram etc.), regression and correlation [10][11][12]. Those methods apply to quantitative or qualitative analysis of data. There was no effort to develop any learning model which is able to learn from the accumulated data and environment factors and predict the most vulnerable geographical location for the epidemic like HIV/AIDS[13][14].

Most of the data analysis in geography has been essentially based on traditional statistics and multidimensional data analysis and does not take into account spatial property [15][16][12]. Some geo-statistical tools are there like MathSoft[4] and spatial analyst for ArcView for ESRI[18] which allows specialized analysis i.e. mapping the statistics analysis result but it lacks the spatial fuzzy classification and correlation feature. The prediction model proposed here is based on ANN model which is far more accurate and flexible than the conventional multi-regression model[19][20][12]. The previous works have been focused on descriptive methods rather than predictive methods[21][22]. This occurs in developing from one hand, spatial statistic methods such as the global and local spatial auto-correlation indices, geographical clustering [23], and from the other hand, SDM methods including generalization and characterization[6]. Besides this an adaptation of the Neural Network based learning model has been proposed that proposes to learn from the spatial and non spatial databases and to improve SDM algorithm.

2. Methodology
2.1 Hypothesis : Migrants, particularly the bridge population, bear a height risk of spreading HIV infection, which results from the condition and structure of the migration process. Available evidence suggests that migration could be fuelling the spread of HIV epidemic in high-out migration states. The analysis of the recent sentinel surveillance data (2008-09) shows that out of the 0.12 million estimated new infections in 2009, the six high prevalence states accounted for only 39% of the cases, while it is accounted for 41% in the states having high out migration rate. Now this HIV prevalence data is combined with the thematic data relating to the location of states. It is hypothesized that location wise Human Development Index [HDI] ,Literacy Ratio female to male[LTFM] ,Average Migration Distance of migrant population [AIMD] and migration rate[MGRT] influences the HIV growth rate [INFDIFF] in a state.

2.2 Model: The Artificial Neural Network [ANN] model used in the study is the multi layer perceptron (MLP). For machine learning, the model requires a desired output which correctly maps input to output. It has got a three layer of architecture (Input, hidden and output layer) as shown in figure 1.

![Figure 1: Back Propagation Artificial Neural Network Architecture](image-url)
Inputs are:  I1: HDI, I2: LTFM, I3: AIMD ,I4: MGRT

Output Y : HIV prevalence [INFDIFF]

Parameters of ANN :
Nodes in input layer: 4
No. of hidden layer: 1
No. of output layer: 1
Nodes in hidden layer: 3
Weight assigned in input to hidden layer: 12 weights
Weight assigned in each hidden to output layer: 3 weights
Learning rate: 0.6

2.3 Algorithm of Backpropagation ANN:

The algorithm of ANN-BPN model is as follows:
The subscripts I_in, H_m, O_q denotes input to input layer, Hidden layer and Output layer neurons. Here n,m,q=1,2,3.... The weight of the architecture between i^th input neuron to j^th hidden layer is I_iW_j and the weight of the architecture between i^th hidden neuron to j^th output layer is H_mW_n.

Input Layer Computation: It is a liner activation function i.e. {O}I={I}I. Here O is the net output of the input layer.

Hidden Layer Computation: The input to the hidden layer(I_H) is the weighted sum of the output of the input neurons. {I}_H=[I_iW_j]^{T}{O}_I. The output of the hidden layer O_H is computed as (sigmoid function) i.e. O_H=\frac{1}{1+e^{-\lambda}}. Here \lambda=0.6

Output Layer Computation: The input to the output layer(I_O) is the weighted sum of the output of the hidden layer. {I}_O=[H_mW_n]^{T}{O}_H. The output of the output layer is also computed as sigmoid function (as mentioned above).

3. Implementation of the Model :

3.1 Database : The relevant data on various state wise non-spatial and spatial attributes have been collected from National AIDS Control Organization[NACO], Planning Commission , UNDP Human Development Report 2009, and Ministry of Statistics and Program Implementation Report October 2011 and stored in ORACLE 10g spatial database.

An overview of state wise data reveals three patterns of HIV growth rate e.g., increasing, stable and decreasing. Accordingly the states with three categories are shown in map of India in different color for different categories [Fig 2].

![Fig 2: The three categories of states having different growth rate of HIV](image-url)
After performing correlation analysis only statistically significant data on four inputs e.g., HDI, LTFM, AIMD and MGRT have been considered for machine learning and other socio economic data which were not so significant are ignored. A graphical plot of the significant input and output data are shown in figures [3 - 5].

3.2 Computational Results:

Data have been filtered using sql query in three categories of states as shown in figures [3-5]. Computer program for running ANN back propagation algorithm has been developed in PL/SQL and stored as a procedure. The SDO_GEOM function in the spatial database of ORACLE 10g computes the average distance of the migrants as one of the significant inputs to the ANN model. Artificial Neural Network is trained with input – output data for each category of states.

For the three categories of states, the mean square [MSE] error graph of the trained ANN model is plotted as shown in figures [6 – 8].
From the plot (Fig. 6) it is evident that the MSE has been minimized up to the level of 0.00044 hence the learning is quite acceptable. The plots of MSE in figures 7 and 8 are 0.00244 and 0.0083 which are again quite low hence acceptable. The converged values of weights between input layer i and the hidden layer j \[ I_iW_j \] and the weights between the hidden layer m and the output layer n \[ H_mW_n \] are shown in table 1 for three types of data sets e.g., with increasing, decreasing and stable HIV growth.

<table>
<thead>
<tr>
<th>Weight s of Nodes</th>
<th>Weight for increasing class</th>
<th>Weight for decreasing class</th>
<th>Weight for stable class</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1W1</td>
<td>.1005744</td>
<td>.1062326</td>
<td>.1245625</td>
</tr>
<tr>
<td>I1W2</td>
<td>.4015816</td>
<td>.4153782</td>
<td>.4738058</td>
</tr>
<tr>
<td>I1W3</td>
<td>.6002843</td>
<td>.6034905</td>
<td>.6108007</td>
</tr>
<tr>
<td>I2W1</td>
<td>-.1264225</td>
<td>-.0582184</td>
<td>-.0946846</td>
</tr>
<tr>
<td>I2W2</td>
<td>.4447993</td>
<td>.5708079</td>
<td>.5212602</td>
</tr>
<tr>
<td>I2W3</td>
<td>-.2763843</td>
<td>-.2251907</td>
<td>-.2530306</td>
</tr>
<tr>
<td>I3W1</td>
<td>.4824482</td>
<td>.56627</td>
<td>.5134122</td>
</tr>
<tr>
<td>I3W2</td>
<td>-.0252463</td>
<td>.1361366</td>
<td>.0465647</td>
</tr>
<tr>
<td>I3W3</td>
<td>.5263539</td>
<td>.5874757</td>
<td>.5506734</td>
</tr>
<tr>
<td>I4W1</td>
<td>.2274652</td>
<td>.246863</td>
<td>.2250029</td>
</tr>
<tr>
<td>I4W2</td>
<td>-.0086566</td>
<td>.0232134</td>
<td>-.0232614</td>
</tr>
<tr>
<td>I4W3</td>
<td>.4088744</td>
<td>.4245377</td>
<td>.4111825</td>
</tr>
<tr>
<td>H1W1</td>
<td>-.7913804</td>
<td>-.1050212</td>
<td>-.8845848</td>
</tr>
<tr>
<td>H2W1</td>
<td>-.14159216</td>
<td>-.1609461</td>
<td>-.15484145</td>
</tr>
<tr>
<td>H3W1</td>
<td>-.5991215</td>
<td>-.8575985</td>
<td>-.718109</td>
</tr>
<tr>
<td>ERR_SQ R</td>
<td>.00044</td>
<td>.00244</td>
<td>.00683</td>
</tr>
</tbody>
</table>

4. Conclusion: This research successfully demonstrates the application of spatial data mining for modeling the prevalence of HIV in India which is found to be significantly influenced by the four parameters e.g., Human Development Index (HDI) ,Literacy Ratio female to male (LTFM) ,Average Migration Distance of migrant population (AIMD) and migration rate (MGRT) .The potential of spatial database of ORACLE 10g has been fully explored in classifying the states on the basis of HIV prevalence rate using SQL query. The machine learning process using back propagation algorithm of Artificial Neural Network has been successfully implemented with PL/SQL procedure developed on ORACLE database.

The mean square error of machine learning process is found to be reasonably low for three types of data sets as defined in section 3.1 above. The computational result corroborates our hypothesis set forth in the section 2.1. This will help the researchers and
planners to work on large spatial database and help the policy makers to plan and monitor the impact of HIV prevention and care intervention program.

5. References