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

Siddhi Nath Rajan
IMS Engineering College, Ghaziabad. (MTU, Noida)
Shobhit University,Meerut

Ashok K Sinha
ABES Engineering College, Ghaziabad, INDIA

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.

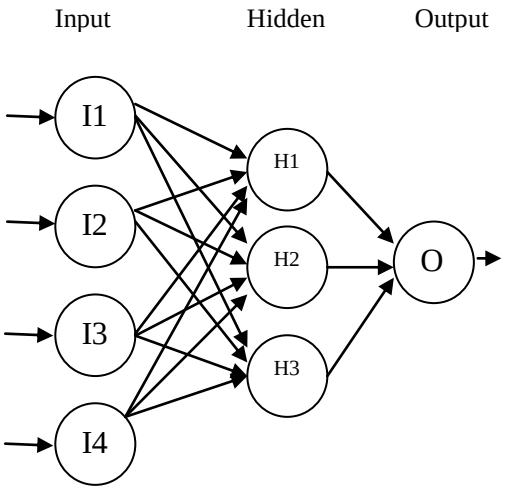


Figure 1: Back Propagation Artificial Neural Network Architecture

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_n,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.

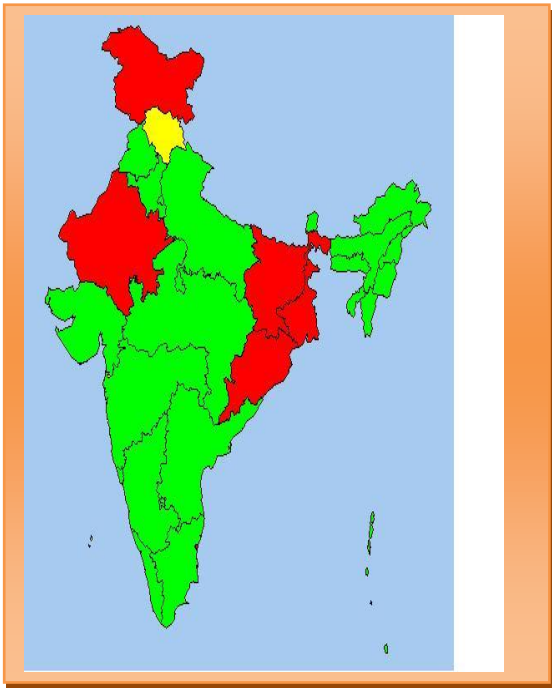
Input Layer Computation: It is a liner activation function i.e. $\{O\}_I=\{I\}_I$. Here O is the net output of the input layer

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. $O_H=\frac{1}{[1+e^{-\lambda(\ I)}]}$. Here λ=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).
Now error $E_r= (T_o-O_o)^2$. Here To is the desired value of the output and O_O is the computed value of the output.

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].



Red Colour: increasing growth rate of HIV prevalence
Green Colour: Decreasing growth rate of HIV prevalence
Yellow Colour: Stable state of HIV prevalence.

Fig 2: The three categories of states having different growth rate of HIV

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].

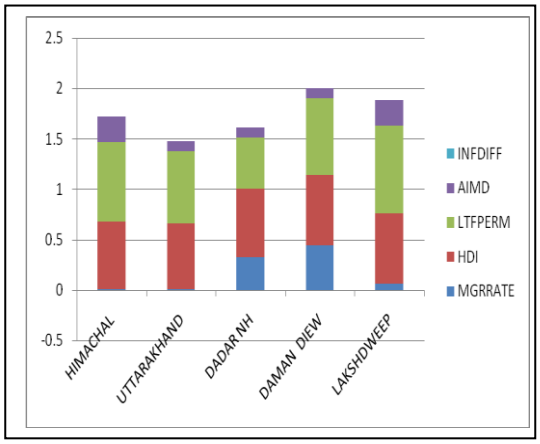


Fig 3: States with stable HIV growth rate

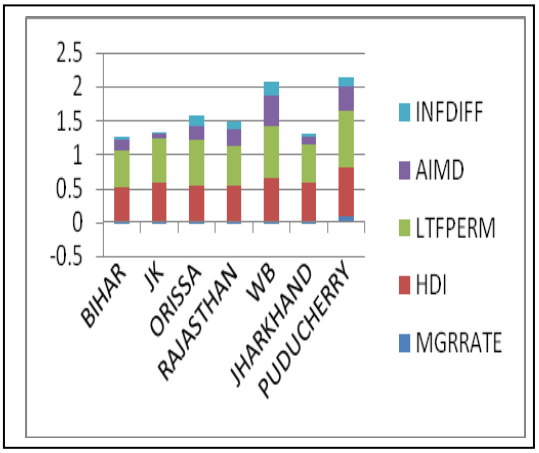


Fig 4: Sates with increasing HIV growth rate

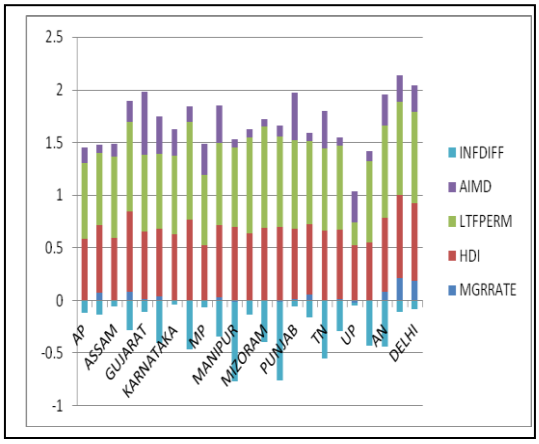


Fig 5: Sates with decreasing HIV growth rate

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].

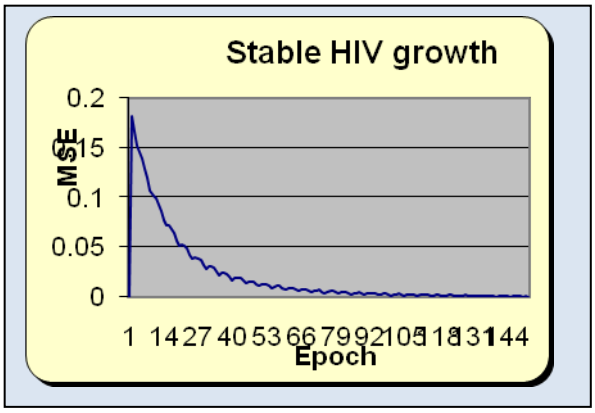


Fig 6: For stable HIV growth states

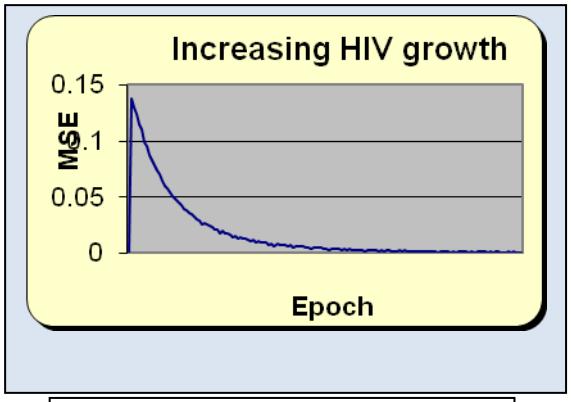


Fig 7: For Increasing HIV growth states

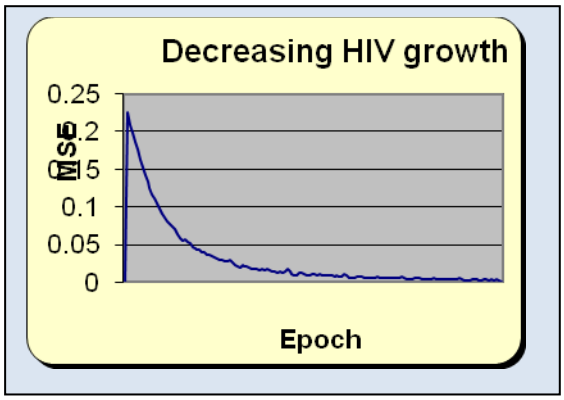


Fig 8: For decreasing HIV growth states

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 [IiWj] and the weights between the hidden layer m and the output layer n [HmWn] are shown in table 1for three types of data sets e.g., with increasing, decreasing and stable HIV growth.

Weights of Nodes	Weight for increasing class	Weight for decreasing class	Weight for stable class
I1W1	.1005744	.1062326	.1245625
I1W2	.4015816	.4153782	.4738058
I1W3	.6002843	.6034905	.6108007
I2W1	-.1264225	-.0582184	-.0946846
I2W2	.4447993	.5708079	.5212602
I2W3	-.2763843	-.2251907	-.2530306
I3W1	.4824482	.56627	.5134122
I3W2	-.0252463	.1361366	.0465647
I3W3	.5263539	.5874757	.5506734
I4W1	.2274652	.246863	.2250029
I4W2	-.0086566	.0232134	-.0232614
I4W3	.4088744	.4245377	.4111825
H1W1	-.7913804	-1.050212	-.8845848
H2W1	-1.4159216	-1.6709461	-1.5484145
H3W1	-.5991215	-.8575985	-.718109
ERR_SQR	**.00044**	**.00244**	**.00083**

Table: 1

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

[1] S. Shekhar and S. Chawla, A Tour of Spatial Databases. Englewood Cliffs, NJ: Prentice-Hall, 2002. ISBN 0-7484-0064-6.
[2] S. Schönfelder. Some notes on space, location and travel behaviour. In Swiss Transport Research Conference, Monte Verita, Ascona, 2001.
[3] Fisher, J.B.; Kelly, M.; Romm, J. Scales of environmental justice: Combining GIS and spatial analysis for air toxics in West Oakland, California. Health Place 2006, 12, 701–714.
[4] U. Ozesmi and W. Mitsch, "A spatial habitat model for the marsh-breeding red-winged blackbird (agelaius phoeniceus 1.)," in Coastal Lake Erie Wetlands Ecological Modeling. Amsterdam, The Netherlands: Elsevier, 1997, vol. 101, pp. 139–152.
[5] R. Pace and R. Barry, "Quick computation of regressions with a spatially autoregressive dependent variable," Geograph. Anal., 1997.
[6]Zeitouni K., "A Survey on Spatial Data Mining Methods Temporal, Spatial and Spatio-Temporal Data Mining, Lecture Databases and Statistics Point of Views", Int. Conf. IRMA, Notes in Artificial Intelligence no2007, Springer,September 12-16,2OOO, Lyon, France, pp102-114.
[7] R. Agrawal, "Tutorial on databse mining," in Proc. 13th ACM Symp. Principles of Databases Systems, Minneapolis, MN, 1994, pp. 75–76.
[8] K. Koperski, J. Adhikary, and J. Han, "Spatial data mining: Progress and challenges," in Proc. Workshop Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada, 1996, pp. 1–10.
[9] S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, and C. T. Lu, "Spatial databases: Accomplishments and research needs," IEEE Trans. Knowledge Data Eng., vol. 11, Jan./Feb. 1999.
[10], "Sparse spatial autoregressions," in Statistics and Probability Letters. Amsterdam, The Netherlands: Elsevier, 1997, pp. 291–297.
[11] J. P. LeSage, "Bayesian estimation of spatial autoregressive models," Int. Reg. Sci. Rev., vol. 20, pp. 113–129, 1997.
[12] R. S. Bivand, E. J. Pebesma, and V. Gomez-Rubio. Applied Spatial Data Analysis with R. Springer Series in Statistics. Springer, New York, 1st edition, 2008.
[13] P. Elliot and D. Wartenberg. Spatial epidemiology: Current approaches and future challenges. Environmental Health Perspectives, 112(9):998{1006, Jun. 2004.
[14] P. Elliott, J.Wake_eld, N. Best, and D. Briggs. Spatial Epidemiology: Methods and Applications. Oxford Medical Publications. Oxford University Press, Oxford, 2nd edition, 2000.
[15] Charre J., Statistique et tenitoire, Editions GIP Reclus, Collection Espaces modes demploi, Montpellier, 1995
[16] Sanders, L., L'analyse statistique des donnbes en ggographie, GIP Reclus, 1989
[17] Mathsoft Inc., "S-Plus for ArcView GIS -Users Guide Version 1.0" and "S-Plus Spatial Stat.", Data Analysis Products Division, Seattle, Washington, April 1998.

[18] Web site of ESRICompany "URL http://www.esri.com"
[19] W. S. Sarle, "Neural networks and statistical models," presented at the 9th Annu. SAS User Group Conf., 1994.
[20]B.P. Vijay Kumar and P. Venkataram. Prediction-based location management using multilayer neural networks. Journal of Indian institute of science, pp.7-21, 2002.
[21] R. Pace and R. Barry, "Quick computation of regressions with a spatially autoregressive dependent variable," Geograph. Anal., 1997.
[22] Jerrett, M.; Burnett, R.T.; Ma, R.J.; Pope, C.A., III; Krewski, D.; Newbold, K.B.; Thurston, G.; Shi, Y.; Finkelstein, N.; Calle, E.E.; Thun, M.J. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology 2005, 16, 727–736.
[23] United States Department of the Interior; United States Geological Survey the Multi-Resolution Land Characteristics (MRLC). Available online: http://www.mrlc.gov/index.asp (accessed February 12, 2009).