ANALELE ŞTIINTIFICE ALE UNIVERSITĂŢII
„ALEXANDRU IOAN CUZA” din IAŞI
Tom LIX, nr.2, s. IIc, Geografie 2013
ISSN 1223-5334 (printed version)

SCIENTIFIC ANNALS OF
„ALEXANDRU IOAN CUZA” UNIVERSITY OF IAŞI
Volume LIX, no.2, s. II c, Geography series 2013
(online version) 2284-6379 eISSN

APPLICATION OF GIS AND STATISTICAL METHODS TO SELECT OPTIMUM MODEL FOR MALARIA SUSCEPTIBILITY ZONATION: A CASE STUDY

Praveen Kumar RAI[1], M.S.NATHAWAT[2]

[1]*Department of Geography, Banaras Hindu University, Varanasi (U.P.), India,* *rai.vns82@gmail.com*
[2]*School of Sciences, IGNOU, New Delhi, India,* *msnatahawat@ignou.ac.in*

Abstract: The representation and analysis of maps of malaria-incidence data is a basic tool in the analysis of regional variation in public health. An attempt has been made for Varanasi district, India to develop malaria susceptibility model using different statistical methods, by which malaria prone zones could be predicted using five classes of malaria susceptibility and comparison of statistical methods to select optimum model for malaria susceptibility zone and verification of the susceptibility zone by area under curve (AUC) though Remote Sensing data and GIS. Multiple linear regression, Information value and heuristic methods are applied for malaria disease occurrence. Using the causal factors and indicators, malaria susceptibility index (MSI) and malaria susceptibility zones (MSZ) are developed. Malaria density ratio (Qs) is used to calculate optimum model for malaria susceptibility index and malaria susceptibility zones. The verification method is performed by comparison of existing malaria data and malaria analysis results by area under curve (AUC). It is found that the information value method having *Qs*=3.96 has been selected as an optimum model for malaria susceptibility zonation in the study area, whereas *Qs* value for Heuristics method and Multiple linear regression method are 1.67 and 1.43 respectively. Verification results show that in the information value case, the area under curve (AUC) is 0.696 and the prediction accuracy is 69.60%. In the heuristic and multiple linear regression case, the AUC is 0.603 and 0.484 and the prediction accuracy is 60.30% and 48.40% respectively.

Key words: *remote sensing, GIS, malaria, NDVI, MSI, MSZ, Qs, AUC.*

I. INTRODUCTION

Malaria is a major health problem in the developing countries and it affects approximate 3.5-5.0 billion people and has overwhelming effects on health and development with at least one million deaths taking place annually. About 70-90 per cent of the risk of malaria is considered due to environmental factors which in turn affect the wealth and survival of the vectors (Smith et al., 1999, Saxena et al., 2009). Malaria remains one of the greatest killers of human beings, particularly in the developing world (Kaya et. al., 2002). Malaria disease transmission depends on various parameters that affect the vectors, parasites, human-hosts etc. These parameters may include, among others, meteorological and environmental conditions (Saxena et al., 2009). The most apparent determinants are observed to be the meteorological and environmental parameters such as rainfall, temperature, humidity and vegetation type and cover (Connor et al. 1997; Craig et al. 1999).

The powerful tools and techniques of the geospatial technologies are very helpful in malaria research. Geospatial technology is the field of information technology that collect, manages, interprets, integrates, displays, manipulate, analyzes and uses datasets concentrating on the geographic, temporal and spatial reference. Geospatial technologies include wide collection of the technologies such as Geographic Information System (GIS), Remote Sensing (RS) and Global Positioning System (GPS). Geospatial technology has clearly defined the epidemiology of disease related to environmental factors by identifying the spatial limits of the disease prevalence and risk mapping with relevant risk factors. The relationships between the disease occurrence and vector distribution could have never been so comprehensively studied without the geospatial technologies. Space born satellite data has made getting information for large and remote areas easy (Saxena et al., 2009).

The representation and analysis of maps of malaria-incidence data is a basic tool in the analysis of regional variation in public health. In the case of disease spread, individuals near or incontact to a contagious person or a tainted environmental setting are deemed more susceptible to certain types of illnesses. Cartographic design and mapping techniques can draw attention to these locations by displaying an aggregation or lack of such events or patterns in space. In this ever increasingly complex world, it is no surprise that the problems that is faced by public health researchers are becoming more and more intricate to solve and resolve. Recently, GIS has emerged as an important component of many projects in public health and epidemiology (Foley, 2002). GIS plays a important roles in the planning and management of the active and complex healthcare facilities system and disease mapping. Although still at an early stage of integration into public healthcare planning, GIS has shown its capability to answer a diverse range of questions relating to the key goals of efficiency, effectiveness, and equity of the provision of public health services (Kleinschmidt et al., 2000; Boscoe et al., 2004; Rai et al., 2012). GIS will play a significant role in the reorganization of public health and disease planning in the twenty-first century in the handling of health information (Srivastava, 2000; Donald et al., 2005; Rai et al., 2012).

Epidemiologists are using spatial map when analyzing associations between location, environment, and disease. GIS has been used in the mapping and monitoring of vector-borne diseases, water-borne diseases, in environmental health, analysis of disease policy and planning, health situation in an area, generation and analysis of research hypotheses, identification of high risk health groups, planning and programming of activities, and monitoring and evaluation of interferences (Rai et al., 2011). GIS-based malaria mapping is using for risk assessment at national, regional, town and village level. Such spatial mapping is considered key for analyzing past as well as present disease trends. Many agencies and government institutions are exploring Health GIS in India and due to sheer size of our country, varied life styles, climatic zones and environmental conditions make it all the more important for India to have a health GIS.

The main aim of the study is to comparison of statistical methods to select optimum model for malaria susceptibility zone and verification of the susceptibility methods by area under curve (AUC) though Remote Sensing data and GIS techniques.

II. THE STUDY AREA

The study area is Varanasi district, U.P., India extending between 25°10′ N to 25°37′ N latitude and 82°39′ E to 83°10′ E longitudes, in eastern Uttar Pradesh, India. Its major portion is stretched towards west and north of the Varanasi city spread over an area of 1454.11 sq. km (Fig.1). The area under study lies in the sub-tropical monsoonal climate which is divided into the following three seasons:

- a cool dry season of northerly winds from October to February.
- a hot dry season from march to mid-June and
- a hot wet season of south-westerly wind from mid-June to September.

In the study area, temperature begins to fall rapidly from November and in January (The coldest month) the maximum temperature comes down to 23.06 degree. The average seasonal temperature varies from 24.65 degree centigrade to 35.05 degree centigrade but when the heat becomes oppressive followed by scorching sun rays and westerly hot winds (locally known as ‘Loo’), mercury often to a maximum of 42.43 degree centigrade. Normally the monsoon bursts in the study area during the third week of June. These hot winds usually cease by mid-June with the advent of south-west monsoon. Some times during this period in association with cold waves, western disturbances bring the minimum temperature upto one or two degree above the freezing point mainly responsible for occurrence of frost.

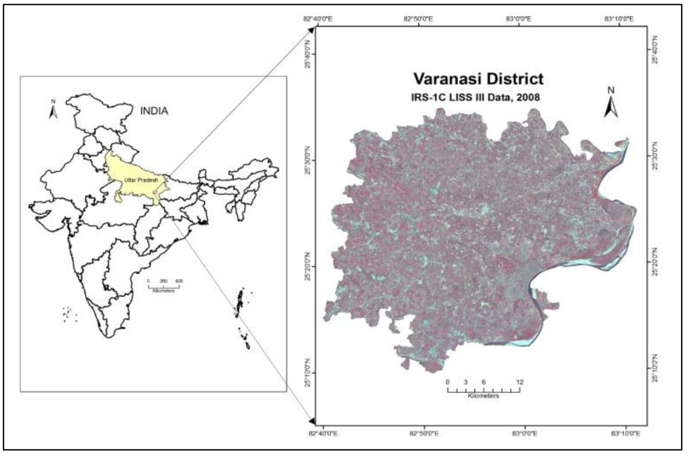

Fig.1. Location of Study Area as Viewed on IRS-1C LISS III Data

The hot wet season of south-westerly wind in the area starts from the third week of June and continued up to the end of September. Decrease in mean temperature up to a minimum of 27.20 degree centigrade and maximum of 30.13 degree centigrade, increase in relative humidity, high precipitation and low monthly range of temperature etc. are the main features of this season. The relative humidity increases from 74.19 % to 79.26 % in the months of July to September month respectively. The October, which is a transitional month between hot rainy and winter seasons, experiences nearly 74 % of high relative humidity but lower than the hot rainy months.

The rainy season in the district generally begins by the end of June and continues till the end of September. This is the most important season for agriculture as it receives nearly 113.47 cm annual rainfall. The moisture content in the soil either by infiltration or capillary rise is fulfilled by rainfall and it is the major ecological factor in agricultural operations starting from ploughing to sowing, germination and ripening of all the crops.

The overall population distribution in the district is closely related to the physical and socio-cultural factors. The Varanasi district, both its urban and rural areas, has a uniquely different growth character, complemented by the movement of people from surrounding areas for occupational reasons, tourist traffic as a result of its heritage value, and special events of spiritual importance of the Ganga at Varanasi. For the present analysis block-wise arithmetic density has been worked out which reveals an irregular and uneven distribution of population density of 997 persons per sq. km. while it varies from 1266 persons per sq. km. in Kashi Vidyapith to lowest of 877 persons per km sq. Km in Pindra Development Block. It is found that the arithmetic density in the study area of year 1991 and 2001, which has been largely influenced by urban centres and terrain conditions.

The study area includes two tahsils namely, Pindra and Varanasi Sadar; which are further sub-divided into eight Development Blocks namely Baragaon, Pindra, Cholapur, Chiraigaon, Harhua, Sevapuri, Araziline and Kashi Vidapeeth, consisting of 1336 villages altogether (Rai et al., 2011, 2012).

III. METHODOLOGY

Important GIS layers on specific parameters which are directly or indirectly related to the occurrence of malaria i.e. land use, Normalized Difference Vegetation Index (NDVI), distance to water ponds, distance to river, distance to road, distance to hospital, rainfall, temperature, projected population density of year 2009, and land use/land cover information are interpreted and produced in remote sensing and GIS platforms using Ilwis Version-3.4 and ArcGIS Version-9.3 and ERDAS Imagine Version-9.1 software. Statistical software SPSS Version-16 is used to produce the layer maps that assist in the production of the malaria susceptibility maps using different statistical methods (Table 1). Survey of India topography (SOI) map of 1:50,000 scale of is used to extract the district and development block boundaries. The reference systems i.e. coordinates and projection of important point for geo reference point like road junctions points and malaria occurrence area, existing health facilities units are delineated and identified during the field visit using Global Position Systems (GPS) technology. The various selected GIS themes are developed from the IRS-1C LISS-III remote sensing data, 2008 and SOI topographical map. Therefore, land use map, NDVI and vector layers of water bodies and other important parameters used in this study are delineated in ERDA Imagine-9.1 and ARC GIS-9.3 software (Rai et. al., 2012).
Three important geo statistical methods are used in this study to produce malaria susceptibility index (MSI) and malaria susceptibility zonation (MSZ) i.e. Multiple Linear Regression, Information Value (Infoval mehtod) and Heuristic Method.

Geo-statistics refers to the collection of statistical methods in which location data plays an important role in the study design or data analysis (Gosoniu et al., 2006, (Saxena et al., 2009). Geo-statistical analysis in malaria disease mapping in relation to exploratory data analysis and disease mapping is discussed in this research. Exploratory data analysis refers to describing patterns in the distribution of a disease using location data (Saxena et al., 2009). Exploratory data analysis for GPS generated location data linked with epidemiological information incorporates statistical methods for point pattern analysis in a GIS (Saxena et al., 2009).

Table 1. Malaria database showing characteristics of malaria based on different parameters

Pameteres for malaria mapping	No. of malaria pixel	Total no. of pixel	Malaria area (%)
A. Rainfall class			
<970	7897	19519	6.38
970-973	19682	68937	15.89
973-976	9542	99487	7.7
976-979	15175	143285	12.25
>984	71547	280162	57.77
B. Temperature class			
35.44-35.46° C	69974	266938	56.5
35.47-35.49° C	24812	234152	20.04
35.50-35.53° C	29057	110300	23.46
C. Population density			
Very Low	41787	222199	33.74
Low	41584	212569	33.58
Moderate	16861	53352	13.61
High	5985	84805	4.83
Very High	17626	38460	14.23
D. Distance to stream			
<1000 m	28766	132325	23.23
1000-3000 m	36354	173300	29.35
3000-6000 m	38783	184054	31.32
6000-10000 m	15533	80756	12.54
>10000 m	4407	40955	3.56
E. Distance to road			
<300m	81541	390258	65.84
300-1000 m	34954	177295	28.22
1000-2000 m	5706	38918	4.61
2000-3000 m	1134	3499	0.92
>3000 m	508	1420	0.41
F. Distance to health facilities			
0-1000 m	8158	29136	6.59
1000-3000 m	21720	145065	17.54
3000-6000 m	43515	242719	35.14
6000-10000 m	39376	147357	31.8
>10000 m	11074	47113	8.94
G. Distance to ponds			
<500 m	67032	270681	50.45
500-1500 m	48608	283869	36.58
1500-3000 m	15909	39612	11.97
>3000 m	1319	17228	0.99
H. NDVI			
-0.288	23909	115299	19.31
0-0986	99934	496053	80.69
I. Land use class			
Water bodies	2690	19282	2.17
Agriculture	65188	330581	52.64
Settlement	25810	114660	20.84
Vegetation	21595	99812	17.44
Fallow land	8544	46601	6.9

Using GIS techniques and GPS device, malaria inventory map is often used as bases for other malaria susceptibility zonation techniques or for an elementary form of a susceptibility map. Village wise malaria locations data are collected from each Primary Health Centers (PHC's) and then locations are determined using GPS devices. Buffer zone or proximity analysis is also used to study the effect of malaria risk factors based interventions and also used to identify disease risk areas where control activities need to be reinforced. GPS location data related to malaria prevalence is imported in GIS platform and 500 m buffer zones are created around each point. On the basis of these malaria pixels falling in the study area, the whole study-area-pixels are assigned two values, i.e., 0 (no malaria pixels) and 1(where, malaria pixels are present).

IV. RESULT AND DISCUSSON

IV.1. Multiple linear regression method (step-wise method)

Multivariate statistical analyses of causative factors controlling malaria disease occurrence may indicate the relative contribution of each of these factors to the degree of disease occurrence within a defined land unit. These analyses are based on the presence or absence of stability phenomena within the units (Van Westen, 1993; Rai et al., 2012). In order to carry out multivariate analysis of data and to determine the all parameters responsible for malaria disease in Varanasi district, a multiple linear regression method is used. Multiple Linear regression models is constructed for malaria cases reported in the study area, as the dependent variable and various time based groupings of temperature, rainfall and NDVI data as the independent variables (Rai et al., 2012). The multiple linear regression method states that how the susceptibility of malaria as the standard deviation of independent variables and interpreters change. All these used parameters are studied in SPSS statistical software using multiple linear regression method and crossed to each other and then finally Malaria Susceptibility Index (MSI) and Malaria Susceptibility Zonation (MSZ) are produced (Rai et al., 2012).

In this study equation of the theoretical model will be described as follows.

$$M = B_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_m X_m + \varepsilon$$

where: M is the existence of malaria in each unit, $X_{1..m}$ are the input independent variables (or instability parameters) observed for each mapping unit, the b's are coefficients estimated from the data through statistical methods, and ε represents the model error (Rai et al., 2012). Using all the above parameters, Malaria Susceptibility Index and Malaria Susceptibility Zones are produced. All these indicators are very helpful to build a relationship between in malaria breeding source.

The following GIS procedures are commonly used in multivariate statistics for malaria susceptibility zonation:

a) Determination of the list of factors that will be included in the analysis. As many input maps are of alphanumeric type, they should be normally converted into numerical maps. These maps can be converted to several maps with presence/absence values for each land unit, or one map with values as percentage cover of each parameter class or expert values according to increasing observed malaria.

b) Convert the observed malaria map into numerical map by attributing the value 1 to observed malaria areas and the value zero to the other areas.

c) Export for each pixel of the map the different numerical values to a statistical package for subsequent analysis.

d) Import the estimated results into a raster map.

e) Classification of the map into susceptibility classes.

In this study, there are ten causative factors for the multiple linear regression models used as independent variables for malaria susceptibility and all variables are numerical. Only one variables i.e. land use is alphanumeric type or categories that cannot be processed immediately.

In order to process the multiple linear regression model of malaria susceptibility with the current data, the SPSS-16 software is used to process the data to estimate the regression.

Table 2. Status of malaria area percentage vs. malaria level based on the multiple linear regression method

Malaria level	Total no. of pixels	Pixel (%)	Malaria area (sq.km)	Malaria area (%)
Very Low	15703	2.57	39.2575	0.36
Low	78946	12.92	197.365	6.66
Moderate	214138	35.03	535.345	27.69
High	220406	36.06	551.015	45.4
Very high	82066	13.43	205.165	19.87
Total	611259	100	1528.1475	100

The malaria model equation used for calculation in MSI and MSZ in multiple linear regression method is:

$$M = (145.54)+(0.009*Pd)+((2.572)*Dh)+((-9.406)*Dwo)+((-1.354)*Ds)+((-1.009)*Dhc+ ((-0.023)*Rf)+ ((-3.47)*At) + ((-1.98)*Dro) + ((-0.106)*NDVI)$$

where M is the occurrence of malaria in each unit.

The Malaria Susceptibility Index (MSI) values from the multiple linear regression method are found to lie in the range from <-15000 to >5000 (Fig. 2).
In the fig. 2 and 3, it is found that 36.06% of the pixel area comes in -4000- 5000 model index class and susceptibility index <-15000 contain 2.57% of the pixels area whereas 13.43% of the pixels area comes in >7000 susceptibility index class. The cumulative frequency curve of MSI values has been segmented into five classes representing near equal distribution to yield five malaria susceptibility zones (MSZ) i.e. very low, low, moderate, high and very high (Table 2, Fig. 2 and Fig. 3). Fig. 3 shows that 45.40% of malaria area comes under high susceptibility level whereas only 0.36% and 6.66% of malaria belongs to very low and low susceptibility level.

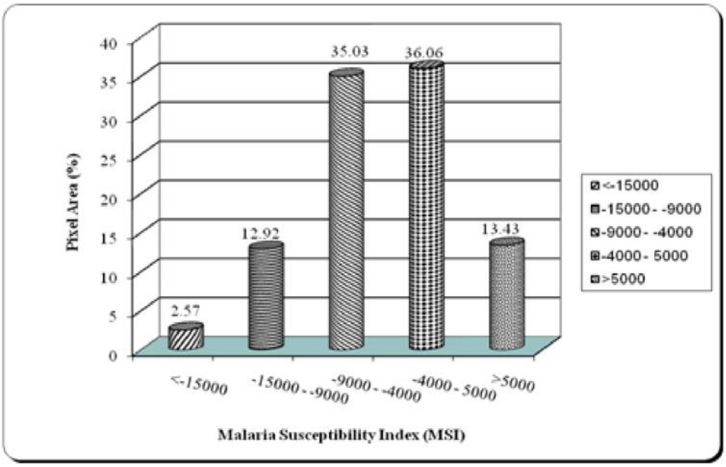

Fig. 2 Distribution frequency histogram of malaria based on the multiple linear regression model.

Among the variables entered in the model based on the standardized Beta values in order of preference are: Pd (with a beta coefficient of 0.211), distance to hospitals, Dh (with a beta coefficient of (with a beta coefficient of 0.196), distance to ponds/water bodies, Dwb (with a beta coefficient of -.154), distance to stream/river, Ds (with a beta coefficient of -.110), distance to health centers, Dhc (with a beta coefficient of -.043), rainfall, Rf (with a beta coefficient of -.215), average temperature, At (with a beta coefficient of -.181), distance to road, Dro (-.020), NDVI (-.015).

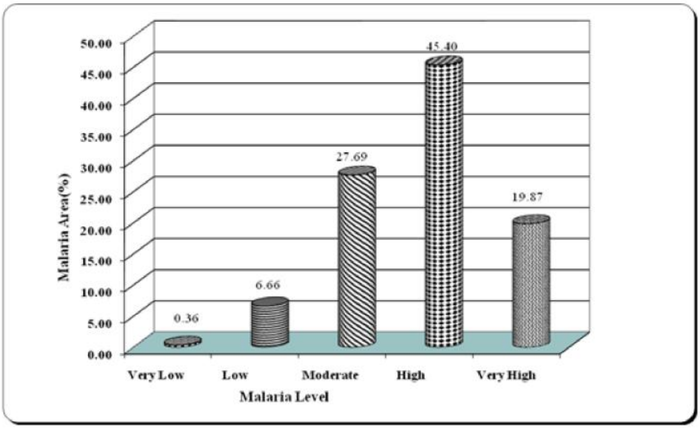

Fig.3 Malaria area vs. malaria susceptibility level based on multiple linear regression method

Among the independent variables, distance from the streams, distance from the road and also because of rainfall, low beta levels are significantly higher than 5% out of the equation (in the statistical calculations done in this manner will be attached to all). Then, calculated coefficients are used in the matrix of dataset and the equation is intended for all of the 38622×9 sample pixels of the study area (Rai et al., 2012). Finally column of the equation result for analyzing and creating malaria susceptibility map has been transferred into GIS software (ILWIS 3.4).

IV.2. Information Value Method (InfoVal*)*

Information Value (InfoVal) method for Malaria Susceptible Zonation considers the probability of malaria occurrence within a certain area of each class of a thematic (Saha et al. 2005). In this model weights of a particular class in a thematic are determined as:

$$W_i = \ln\left(\frac{Densclas}{Densmap}\right) = \ln \frac{Npix(S_i)/Npix(N_i)}{\sum_{i=1}^{n} Npix(S_i)/\sum_{i=1}^{n} Npix(N_i)} \quad \text{(Eq.1)}$$

where W_i is the weight given to the *i*th class of a particular thematic layer, *Densclas* is the malaria density within the thematic class, *Densmap* is the malaria density within the entire thematic layer, *Npix(S$_i$)* is the number of malaria pixels in a certain thematic class, *Npix(N$_i$)* is the total number of pixels in a certain thematic class, and *N* is the number of classes in a thematic map. The natural logarithm is used to take care of the large variation in the weights. Thus, the weight is calculated for various classes in each thematic. The thematic maps are then overlaid and added to prepare a Malaria Susceptible Index (MSI) map. Near-equal subdivision of MSI, cumulative frequency curve categorized into five zones based on malaria susceptibility (i.e., very high, high, moderate, low and very low). Information analyzing includes two specific steps i.e. Bivariate analyze and Multivariate analyze.

IV.2.1. Bivariate Analyze

Determining relationship between affective parameters and their sub-group with malaria is calculated by appraising relative area occupied by malaria in the study area and the area of any sub-group. In this analyze, two ratios are used:

$$P = \frac{M}{N} \textit{ (Eq.2)}, \text{ where:}$$

P is the ratio of malaria area and study area
M is the number of malaria area in the study area and
N is whole of the study area.

$$P_i = \frac{Mi}{Ni}, \textit{ (Eq.3)}, \text{ where:}$$

Pi is the ratio of malaria area in the individual parameters:
*M*i is the number (area) of the malaria area into the *i*th variable
Ni is the whole of the study area including *i* th variable. Then the relation between these two values entitled information value resulted from *i*th variable in prediction malaria potential characterized with *Mi*.

$$Mi = \frac{Pi}{P} = \frac{\frac{Mi}{Ni}}{\frac{M}{N}} \textit{ (Eq.4)}$$

Then, for each variable and its sub-group the Mn of each Mi, which made +ve and –ve zones, has been calculated. If calculated *Mn*s are +ve indicated that the pixels including *i* variable has malaria more than mean of the study area. This indicates the susceptibility of these parameters to instability. While –ve values show the stability means no presence of malaria pixels.

IV.2.2. Multivariate Analyze

After producing thematic maps by interpolating results of each variable information value, by cutting the maps in 200×200 pixels, samples have been apply. Then the numerical values results, which including 38622 samples have been transferred to (EXCEL) software and using the *(Eq.5)* final information value has been calculated.

$$Ij = \quad x_{ji} * Ii = \quad x_{ji} * \frac{\frac{Mi}{Ni}}{\frac{M}{N}} \textit{ (Eq.5)}$$

with (j=1, 2, 3… n) indicate number (area) of networks, (*i*=1, 2… n) indicate the number of variables, Xj*i* is the quantity of i*th* variable in the j*th* indicator, if *i* variable present the value=1 (Malaria), otherwise value=0 (No Malaria) and *Ii* is the information values result from *i*th variable.

After statistical calculation of the model, information resulted from model has been transferred to GIS (ILWIS 3.4) and the MSI map has been created (Fig. 4). The next step in this method is determining the quantities of crucial information values and dividing the degree of susceptibility, which are based on calculated values. For the malaria, the crucial quantity can be definite as the quantity by which the frequency of malaria in the higher quantities is high. To calculate this amount, the information values-calculated from malaria area, distribution frequency histogram has been used (Fig. 4).

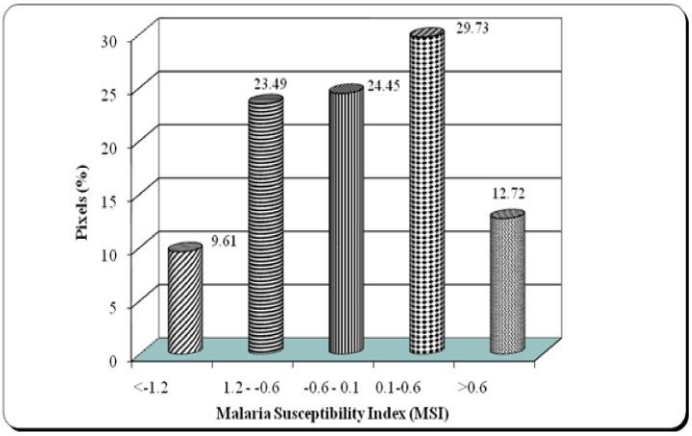

Fig. 4 Distribution frequency histogram of malaria based on the information value model

As shown in this histogram, distribution of malaria area in the information value=0.1-0.6 is looking sensitive, 29.73% of pixels area have the quantities more than this amount, so this value can be definite as crucial value for malaria. Pixels networks having information value more than 0.1 based on malaria area percentage have been divided into two groups i.e. high and very high susceptible and pixels having lower than 0.1 information value have been divided into three level, low, very low and moderate susceptibility (Table 3 and Fig. 5). Table 3 and fig. 5 also highlighted that, 3.87% of the malaria area comes under very low malaria susceptibility level whereas 39.86% and 26.29% of the malaria area falls in high and very susceptibility level respectively.

Table 3. Status of malaria area percentage vs. malaria level based on the information value method (InfoVal)

Malaria level	Total no. of pixels	Pixel (%)	Malaria area (sq.km)	Malaria area (%)
Very Low	58782	9.61	47.95	3.87
Low	143609	23.49	149.61	12.08
Moderate	149494	24.45	221.53	17.89
High	181759	29.73	493.7	39.86
Very High	77746	12.72	325.64	26.29
Total	611390	100	1238.43	100

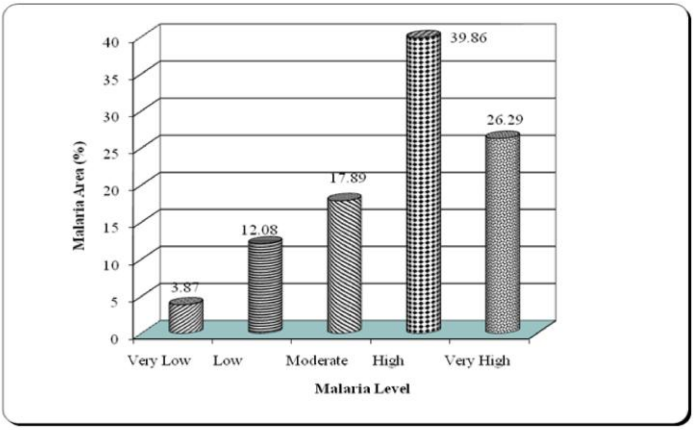

Fig. 5 Malaria area vs. malaria susceptibility based on the information value method

IV.3. Heuristic Approach (Qualitative Map Combination)

The heuristic approach or weighting method is based on the expert opinion and the relative importance of various causative parameters derived from field knowledge. The various data layers as distance to hospitals, distance to health centers, distance to streams, distance to ponds, rainfall etc. have been arranged in a weighting values (from 1 to 5) and similarly, each class within a layer has given a weighting values, the highest class has 5 value, the medium class 3 value and the lowest class has 1 value.

The weights are assigned to the classes of each thematic layer respectively, to produce weighted thematic maps, which have been overlaid and numerically added according to *(Eq.6)* to produce a Malaria Susceptibility Index (MSI) map.

$$MSI = Pd + Rf + Dri + Dpo + Dhf + Dro + Temp + Lu/Lc + NDVI \qquad (Eq.6)$$

where: *Pd, Rf, Dri, Dpo, Dhf, Dro, Temp., Lu/Lc and NDVI* are distribution-derived weights for Population density, Rainfall, Distance to river/streams, Distance to ponds, Distance to health facilities, Distance to road, Temperature, Land Cover and NDVI respectively.

The MSI values from the Weighting method are found to lie in the range from 21 to 37 and (Fig. 6 and 7). It is cleared from the fig. 6 that model index value 21-24 consist 1.71% of the malaria pixels whereas 51.79% and 14.35% of the pixels comes in 27-30 and 33-37 model index classes respectively. The cumulative frequency curve of Malaria Susceptibility Index (MSI) values are divided into five classes shows five malaria susceptibility zones i.e. very low, low, moderate, high and very high (Fig.7).

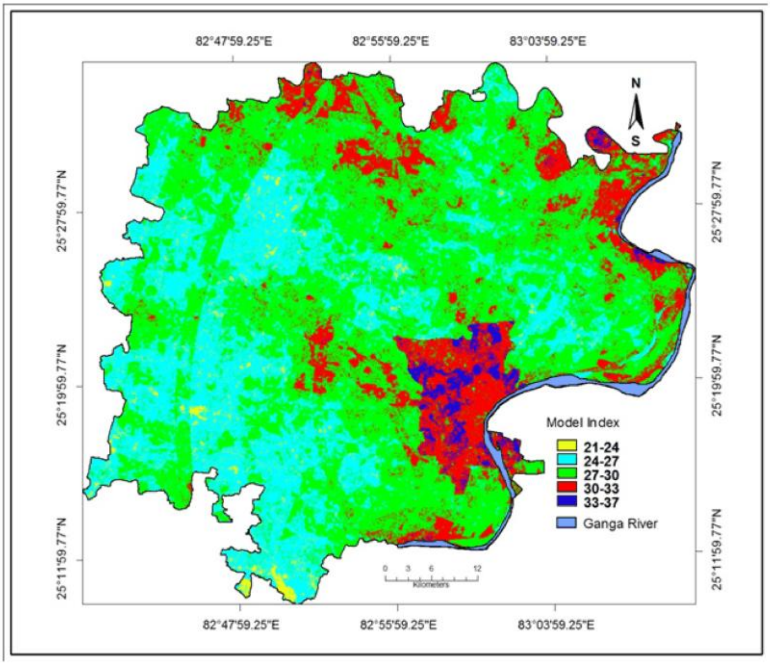

Fig. 6 Malaria Susceptibility Index (MSI) through heuristic approach (Qualitative map combination)

In the table 4, it is shown that 51.78% of malaria area comes in moderate class where as 0.94% and 3.98% of the malaria area comes in very low and very high classes respectively.

Table 4 Status of malaria area percentage vs. malaria level based on the heuristic approach (qualitative map combination)

Malaria level	Total no. of pixels	Pixel (%)	Malaria area (sq.km)	Malaria area (%)
Very Low	10448	1.71	11.68	0.94
Low	183170	29.98	265.46	21.43
Moderate	316394	51.79	641.22	51.78
High	87664	14.35	270.62	21.85
Very High	13201	2.16	49.29	3.98
Total	611390	100	1238.43	100

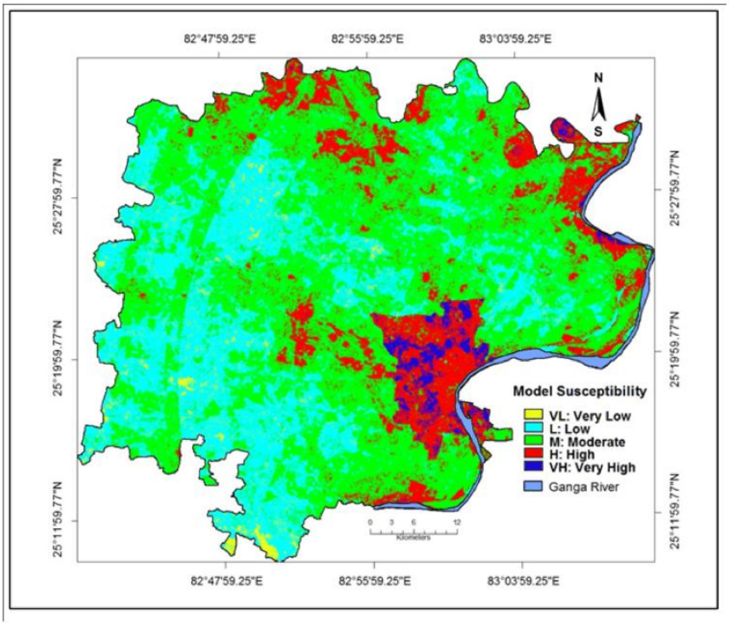

Fig. 7 Malaria Susceptibility Zone (MSZ) through heuristic approach

IV.4. Comparison of statistical methods to select optimum model for MSZ by Qs method

The second aim of this study are choosing optimum model for malaria susceptibility zonation. For appraisal the results of the models, malaria density ratio (*Qs*) method has been used. To appraisal the results in this method following formula (Eq.7) has been used:

$$D_r = \frac{d\prime}{d} = \frac{Sl\prime}{Sa\prime}\Big/\frac{Sl}{Sa} \quad (Eq.7)$$

Where: D_r is the ratio of whole malaria, *d'* is the density of malaria in the susceptibility level, *d* is the malaria density in the area, *Sl'* is the malaria area pixels in special susceptible level, *Sl* is the area of special susceptible level, *Sa'* is the malaria area pixels in the study area and Sa is total pixels in the study area.

$$Qs = \ {}_{i}^{n}(D_r - 1)\ \ D_r - 1\ \ ^{2}x\%S\ (Eq.8)$$

Where: *S* is the ratio of the area of each susceptible level of the study area and *Qs* is the quality score of each models. Based on this equation the results have been shown in the table 5 and fig. 8, which shows the *Dr* Values for each models of malaria susceptibility class.

As shown in the table 5 and fig. 8, it is found that the information value method having *Qs*=3.96 has been selected as optimum model for malaria susceptibility zonation in the study area, whereas *Qs* value for Qualitative map combination (Heuristics method) and Multiple linear regression method are 1.67 and 1.43 respectively.

Table 5 Results of the Qs for the Used Models

Susceptibility level Method	Very low	Low	Moderate	High	Very high	Qs
Information value	0.4	0.41	0.58	1.06	1.63	3.96
Qualitative map combination	0.55	0.57	0.79	1.2	1.46	1.67
Multiple linear regression	0.14	0.41	0.62	1	1.17	1.43

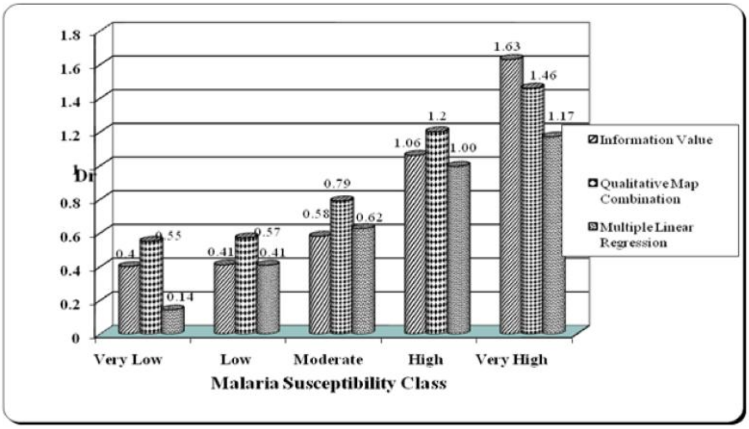

Fig. 8 Dr Values of Qs Method vs. Malaria Susceptibility Class.

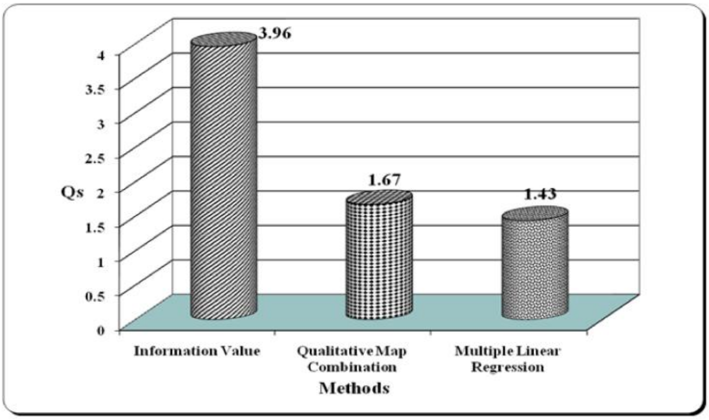

Fig.9 Results of the Qs for the Used Models.

IV. 5 Susceptibility methods verification by area under curve (AUC)

The malaria analysis is performed using the information value, multiple linear regressions, and heuristic models, and the analysis results are verified using the malaria locations for the study area; the result of this is shown in the Fig. 10. The verification method is performed by comparison of existing malaria data and malaria analysis results. The comparison results shown in Table 6 and as a line graph in the Fig. 10 illustrate that how well the estimators perform with respect to the malarias used in making those parameters. Verification of the model success rate is based on the malaria susceptibility analysis result in Varanasi district using the malaria occurrence locations for the three analysis methods i.e. information value, multiple linear regression, and heuristic models. The rate curves are created and the ‘‘areas under the curves’’ are calculated for all three cases of susceptibility maps using the existing malaria location data. The ‘‘areas under the curves’’ constitutes one of the most commonly used accuracy statistics for the prediction models in environmental assessments (Begueria, 2006).

The rate explains how well the model and factor predict the malaria (Chung, 2004). So, the area under the curve can be used to assess the prediction accuracy qualitatively. To obtain the relative ranks for each prediction pattern, the calculated index values of all cells in the study area are sorted in descending order. The ordered cell values are then divided into 100 classes and set on the y-axis, with accumulated 1% intervals on the x-axis. The rate verification results appear as a line in the Fig. 10 and Table 6. For example, in the case of the information value model used, 90– 100% (10%) class of the study area where the malaria index had a higher rank could explain 28% of all the malaria in the success rate and is classified as ‘‘very highly susceptible’’ zone. The next 80–100% (20%) class of the study area where the malaria index had a higher rank could explain 48% of the malarias in the success rate and is classified as ‘‘high susceptible’’ zone. Similarly, the 60–100% (40%) class of the study area where the malaria index had a relatively lower rank could explain 69% of the malaria in the success rate and is classified as ‘‘moderately susceptible’’ zone.

Table 6 Verification and success rate for the study area

Range	Success rate curve (Information value)	Success rate curve (heuristic)	Success rate curve (multiple linear regression)
100–100	0	0	0
90-100	28	20	6
80-100	48	34	14
70-100	60	45	21
60-100	69	54	32
50-100	76	65	46
40-100	82	73	50
30-100	88	81	79
20-100	94	88	87
10-100	96	94	93
0-100	100	100	100

Finally, the remaining 40–100% (60%) class of the study area where the malaria index had a low rank could explain 82% of the malarias are classified as ‘‘not susceptible’’ zone. The same procedure is adopted for classification and verification of the hazard maps obtained through heuristic (Table 6, column 2) and multiple linear regression (Table 6, column 3) models. To compare the result quantitatively, the areas under the curve are re-calculated as the total area is 1 which means perfect prediction accuracy. So, the area under a curve is used to assess the prediction accuracy qualitatively, as shown in the Fig. 10. From the Table 7 and Fig. 10, verification results show that in the information value case, the area under curve (AUC) is 0.696 and the prediction accuracy is 69.60%. In the heuristic case, the AUC is 0.603 and the prediction accuracy is 60.30%. In the multiple linear regression case, the AUC is 0.484 and the prediction accuracy is 48.40%. So from the success rate graphs (Fig. 10), it is quite evident that, Information Value has the best prediction accuracy of 69.60%, whereas the multiple linear regression has the worst accuracy of 48.40%, with a difference of about 21.2%.

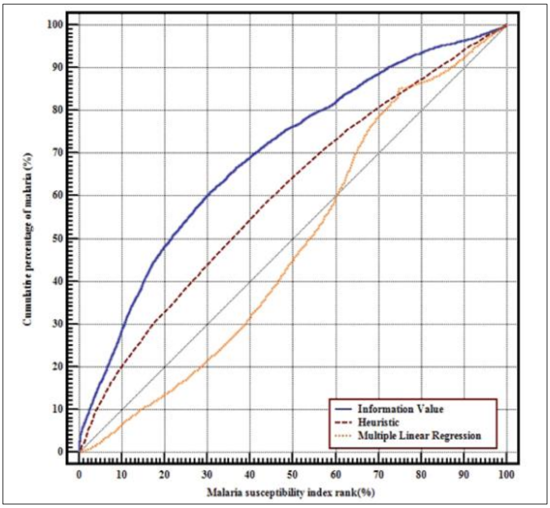

Fig. 10 Illustration of cumulative Frequency Diagram Showing Malaria Susceptibility Index Rank (x-axis) Occurring in Cumulative Percentage of Malaria Occurrence (y-axis).

Although, for the first seven classes (30–100%), the heuristic model is better than from the multiple linear regression model, except for the remainder of the classes (70–100%), the heuristic model produced somewhat similar results to those from the information value and multiple linear regression models. The multiple linear regression model is worse than the information value and heuristic models in all classes.

Table 7 Verification Results Using Area under Curve (AUC)

Method	AUC	Prediction accuracy (%)	Standard error (SE)	95% confidence interval (CI)
Information Value	0.696	69.6	0.00357	0.691 to 0.701
Heuristic	0.603	60.3	0.00371	0.598 to 0.608
Multiple linear regression	0.484	48.4	0.00368	0.479 to 0.489

V. CONCLUSION

Accurate calculation of malaria disease risk is dependent on knowledge of a number of parameters i.e., Land Use, NDVI, climatic factors, distance to location of existing government health centers, population, distance to ponds, streams and roads etc., that are related to malaria transmission. Climatic factors i.e. rainfall, temperature and relative humidity are known to have a strong influence on the biology of mosquitoes for malaria disease. All these parameters are analysed and calculated in SPSS statistical software using three statistical methods i.e. multiple line regression method, information value (infoval) and heuristic approach and crossed to each other and then finally malaria susceptibility index (MSI) and malaria susceptibility zonation (MSZ) are produced. The second aim of this study is choosing optimum model for malaria susceptibility zonation. For appraisal on the results of the models, malaria density ratio (Qs) method has been used and it is found that it is found that the information value method having Qs=3.96 has been selected as optimum model for malaria susceptibility zonation in the study area, whereas Qs value for qualitative map combination (heuristics method) and multiple linear regression method are 1.67 and 1.43 respectively. The verification method is performed by comparison of existing malaria data and malaria analysis results. The area under curve (AUC) is used to assess the prediction accuracy qualitatively and verification results show that in the information value case, the area under curve (AUC) is 0.696 and the prediction accuracy is 69.60%. In the heuristic case, the AUC is 0.603 and the prediction accuracy is 60.30%. In the multiple linear regression case, the AUC is 0.484 and the prediction accuracy is 48.40%. It is quite evident that, Information Value has the best prediction accuracy of 69.60%, whereas the multiple linear regression has the worst accuracy of 48.40%, with a difference of about 21.2%.

Acknowledgements: The first author expresses gratitude to Mr. Mohhamad Onagh Research Scholar, Department of Geography, B.H.U., Varanasi and Mr. M.A. Khan, District Malaria Officer, Varanasi, India for supporting the data collections.

References

Begueria S.: Validation and evaluation of predictive models in hazard assessment and risk management, Natural Hazards., 37,315–329, 2006.

Boscoe F.P., Ward M.H., Reynolds P.: Current practices in spatial analysis of cancer data: data characteristics and data sources for geographic studies of cancer. International Journal of Health Geography, 97: 14041-3, 2004.

Chung K., Yang D.H. and Bell R.: Health and GIS: Toward spatial statistical analyses, Journal of Medical Systems., Vol. 28, No. 4, 349-360, 2004.

Connor S.J., Flasse S.P, Perryman A.H. and Thomson M.C.: The contribution of satellite derived information to malaria stratification, monitoring and early warning. World Health Organization mimeographed series. WHO/MAL, 1079, 1997.

Craig M.H., Snow R.W. and Le Sueur D.: A climate-based distribution model of malaria transmission in sub-Saharan Africa. Parasitol Today., 15, 105-111, 1999.

Donald P.A., Wilbert M.G., Barbara L.: Spatial analysis, GIS and Remote Sensing application in the health, Ann Arbor Press, Chelsea, Michigan, 185, 2006.

Foley, R.: Assessing the applicability of GIS in a health and social care setting: planning services for informal cares in East Sussex, England. Social Science and Medicine, 55, 79–96, 2002.

Gatrell A. and Loytonen M.: GIS and health. London: Taylor & Francis, 1998.

Gosoniu L., Vounatsou P., Sogoba N. and Smith T. Bayesian modeling of geostatistical malaria risk data, Geospatial Health, 1:127-139, 2006.

Kaya S., Pultz T.J., Mbogo C.M., Beier J.C., and Mushinzimana E: The use of radar remote sensing for identifying environmental factors associated with malaria risk in coastal Kenya. International Geoscience and Remote Sensing Symposium, Toronto, 2002.

Kleinschmidt I., Bagayoko M., Clarke G.P.Y., Craig M. and le Sueur D.: A spatial statistical approach to malaria mapping. International Journal of Epidemiology, 29: 355-361, 2000.

Messina J.P. and Crews-Meyer K.A.: The Integration of remote sensing and medical geography: process and application. In: Donald P.A. et al. (eds.): Spatial Analysis, GIS, and Remote Sensing Applications in the Health Sciences Ann Arbor Press, Michigan, 156, 2005.

Mullner R.M., Kyusuk C., Croke K.G. and Menash E.K.: Geographical information systems in public health and medicine. Journal of Medical System; 28, 3; 215-221, 2004.

Omumbo J.A., Hay S.I., Snow R.W., Tatem A.J. and Rogers D.J.: Modelling malaria risk in East Africa at high spatial resolution. Tropical Medicine and Internal Health, 10, 6: 557-566, 2005.

Rai P.K., Nathawat M.S. and Onagh M.: Application of multiple linear regression model through GIS & Remote Sensing for malaria mapping in Varanasi district, India. Health Science Journal, 6, 4: 731-749, 2012.

Rai P.K., Nathawat M.S., Mishra A., Singh S.B. and Onagh M.: Role of GIS and GPS in VBD Mapping: A Case Study. Journal of GIS Trends. Academy Science Journals, 2(1): 20-27, 2011.

Riedel N., Vounatsou P., Miller J.M., Gosoniu L., Chizema-Kawesha E., Mukonka V., Steketee R.W.: Geographical patterns and predictors of malaria risk in Zambia: Bayesian geostatistical modeling of the 2006 Zambia national malaria indicator survey (ZMIS). Malaria Journal, 9, 37: 2010.

Rytkönen Mika J.P.: Not all maps are equal: GIS and spatial analysis in epidemiology. International Journal of Circumpolar Health, 63: 9-24, 2004.

Saha A.K., Gupta R.P. and Arora M.K.: GIS based landslide hazard zonation in Bhagirathi, Ganga valley, Himalayas. International Journal of Remote Sensing, 23: 357-369, 2005.

Saxena R., Nagpal B.N., Srivastava A., Gupta S.K. and Dash A.P.: Application of spatial technology in malaria research and control: some new insights, Indian Journal of Medical Research, 125-132, 2009.

Smith K.R., Corvalán C.F., Kjellström T.: How much global ill health is attributable to environmental factors? *Epidemiology*; *10,* 573-84, 1999.

Srivastava A. and Nagpal B.N.: Mapping malaria. GIS development, IV (6), 28-31, 2000.

Sudhakar S., Srinivas T., Palit A., Kar S.K. and Battacharya S.K.: Mapping of risk prone areas of kala-azar (Visceral leishmaniasis) in parts of Bihar state, India: an RS and GIS approach. Journal of Vector Borne Disease, 43:115–122, 2006.

Sweeney A.W.: A spatial analysis of mosquito distribution. GIS Use, 21: 20-21, 1997.

Van Westen CJ.: Application of Geographic Information Systems to Landslide Hazard Zonation, PhD dissertation, Technical University Delft, ITC publication no. 15

Received: 25.10.2013
Revised: 05.12.2013
Accepted: 11.12.2013
Published: 26.12.2013