Lund University GEM thesis series nr 13

A Comparison of Four Methods of Diseases Mapping

Chao Yang

2016
Department of Physical Geography and Ecosystem Science
Lund University
Sölvegatan 12
S-223 62 Lund
Sweden

LUND UNIVERSITY

UNIVERSITY OF TWENTE.

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

A Comparison of Four Methods of Diseases Mapping

By

Chao Yang

Thesis submitted to the department of Physical Geography and Ecosystem Science, Lund University, in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation for Environmental Modelling and Management

Thesis assessment Board

First Supervisor: Dr. Ali Mansourian (Lund University)
Co-supervisors: Mohammadreza Rajabi (Lund University)

Exam committee:
Examiner 1: Lars Harrie (Lund University)
Examiner 2: Per-Ola Olsson (Lund University)

Disclaimer

This document describes work undertaken as part of a program of study at the University of Lund. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

Course title: Geo-information Science and Earth Observation for Environmental Modelling and Management (GEM)

Level: Master of Science (MSc)

Course duration: January 2016 until June 2016

Consortium partners:

The GEM master program is a cooperation of departments at 5 different universities:
University of Twente, ITC (The Netherlands)
University of Lund (Sweden)
University of Southampton (UK)
University of Warsaw (Poland)
University of Iceland (Iceland)

Abstract

Disease susceptibility mapping can produce risk maps showing predictive distribution of disease incidences. Hence, it is a useful tool for disease prevention. Data-driven approaches for disease susceptibility mapping model relationships and patterns from experience data. Therefore, using data-driven methods can facilitate a simple and direct approach towards disease susceptibility mapping.

In this study, four data-driven models: logistic regression (LR), backpropagation neural network (BPNN), radical basis functional link nets (RBFLN) and general regression neural network (GRNN) for disease susceptibility mapping were implemented using Python. The performances of those four models were tested by a case study of visceral leishmaniosis (VL). Seven VL occurrence related factors, which are temperature, proximity to river, precipitation, proximity to nomadic villages, land cover, altitude, proximity to health-centers, were fed into the models as the input data.

In the results, the area under the receive operating characteristic curve (AUC) was used to test the discrimination ability of the models, and the BPNN generated the best AUC of 0.942. The GRNN classified 70% of the validation data into the right class. The AUC generated by RBFLN, LR and GRNN are 0.938, 0.899 and 0.88 respectively. Meanwhile, BPNN, RBFLN and LR correctly classified 80% of the validation data. All of them they predict that the northern and south-eastern part of the study area have high susceptibility in VL incidence.

Content

List of Figures

List of Tables

List of Equations

1. Introduction

1.1. Background

Disease prevention is an effective method to dramatically minimize fatalities from diseases. According to WHO (WHO 2016a), vector-borne diseases account for more than 17% of all infectious diseases and cause more than a million deaths annually, even if many of these diseases are preventable through informed protective measures.

Disease susceptibility mapping using GIS (geographical information science) can be a useful tool for disease prevention. Risk maps from disease susceptibility mapping can show prevalence distribution and can highlight areas prone to suffer certain endemic hazards due to their prevailing social and environmental conditions (Leonardo et al. 2007). Disease susceptibility mapping can produce risk maps showing predictive distribution of disease incidences. It can help disease managers make better planning of control activities and facilitate the establishment of disease early warning systems that can contribute to reducing the negative socioeconomic effect of the disease.

Studies have shown that susceptibility maps can be produced by modeling association between incidences of diseases and environmental/demography factors (Salahi-Moghaddam et al. 2010; Tsegaw et al. 2013; Seid et al. 2014). Knowledge-driven and data-driven are two possible techniques for disease mapping. Knowledge-driven models rely on the expertise of analysts to assign weights to a series of factors, which can be difficult without relevant expert knowledge. Data-driven approach model relationships and patterns from the experience data. Data-driven methods despite of knowledge-driven methods provide a simpler and more direct approach for disease susceptibility mapping.

There are a variety of data-driven methods conducted (Yesilnacar and Topal 2005; Lee and Sambath 2006; Rajabi et al. 2014). Among the current methods logistic regression (LR) and artificial neural networks (ANNs) are two popular ones. LR is a multivariate statistical method which can be used to predict the probability of occurrence or absence of a certain situation. It is a widely used statistical method

which has a simple structure but can generate good result in making risk maps (Lee and Sambath 2006; Yilmaz 2009). ANNs are extensively used to handle multi-dimensional non-linear characteristics of many real-life problems (Fung et al. 2005). However, few studies have implemented such methods in generating disease susceptibility maps.

1.2. Problem statement

There are many methods that can be used for disease susceptibility mapping. However, the performance of these methods for disease susceptibility mapping are rarely investigated.

1.3. Aim

The main objective of this thesis is to develop and compare four methods for disease susceptibility mapping, namely: logistic regression (LR), backpropagation neural network (BPNN), radical basis functional link nets (RBFLN) and general regression neural network (GRNN). To achieve this, the following objectives should be met:

- Implement four models (logistic regression, BPNN, RBFNN and GRNN) of making disease susceptibility map in Python.
- Compare and test those four methods on visceral leishmaniosis (VL) as a case study disease.

1.4. Methodology

The methodology followed is focused on 2 main outputs: The first one is to implement the four approaches in disease susceptibility mapping, and then do a comparative analysis on results.

Figure 1.1 shows the process describing the activities conducted. An intensive review of past studies and literature about using LR and ANNS on susceptibility mapping was undertaken. Then, four methods were chosen for further analysis. Theoretical and implementation knowledge about the four selected methods were also analyzed. After that, the models of susceptibility mapping were implemented using Python and tested by one case study of VL. The results generated from the VL case study were then analyzed based on the aspects of discrimination and calibration from each of these models. Finally, generalizations and conclusions regarding the advantages and drawbacks of each model were made.

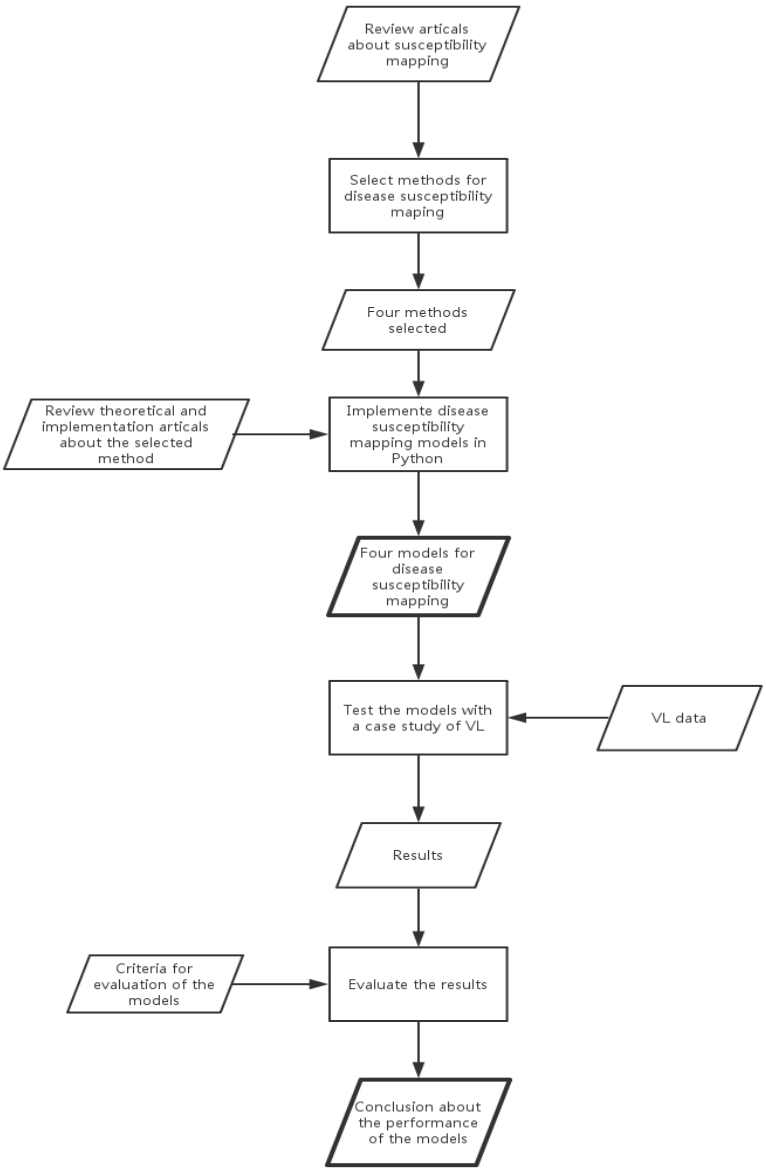


Figure 1-1.Workflow of comparing four models of disease susceptibility mapping.

1.5. Outline of the thesis

This thesis is structured in five main chapters. Chapter 1, Introduction, describes the background information as well as the general outline of the thesis. Chapter 2, Literature review, describes the algorithms compared in the thesis and the using of those algorithms in GIS. Chapter 3, Case study of VL, describes the implementation of those algorithms in a case study of VL. Chapter 4, Result, presents the results generated by the case study. The analysis of the comparisons of the results is discussed in Chapter 5, Discussion. Chapter 6, Conclusion, describes all the work done and the summary of the results and discussion.

2. Literature review

This study is a comparison of four methods in making disease susceptibility mapping. It is constituted into two main parts: the technique and the application of those methods in susceptibility mapping. To cover both, a review of important related literature and previous works are discussed in this chapter.

This chapter is structured into four main parts: section 2.1 focuses on the review of the logistic regression; section 2.2 discusses three selected types of ANNs; section 2.3 introduces the visceral leishmaniosis, which was used in the case study to test the performance of the techniques; and, section 2.4 is the review of the application of those methods in GIS, especially in susceptibility mapping.

2.1. Logistic regression

Regression can be considered as a process to find a series of coefficients which can best describe the relation between the observation values and the dependent values. LR is a commonly used regression approach. It is a statistical method, which is usually used to find the relationship between several independent variables and the probability of a binary or categorical response (Lee and Sambath 2006). The advantage of logistic regression is that, by adding a proper link function to the normal linear regression model, the variables can be any combination of continuous and discrete, and they do not have to be normal distribution (Lee and Sambath 2006). As the response value is the probability of absence or presence of an event, it can be used to predict the probability of the disease incidence via a series of social and environmental factors. The relation between the occurrence and the independent variables can be expressed as:

$$P = 1 / (1 + e^{-z}) \qquad \text{Equation 1}$$

$$Z = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n \qquad \text{Equation 2}$$

Where P is the probability of the occurrence of an event, and Z is a linear combination of the independent variables, for example an n-dimensional social-environmental factors. The b_0 is the intercept of the model and b_1, $b_2 \ldots b_n$ are the weights of the factors.

2.2. Artificial neural networks

In machine learning and cognitive science, artificial neural networks are a family of models inspired by the sophisticated functionality of human brain where hundreds of billions of interconnected neurons process information in parallel. They are widely used to estimate or approximate functions that can predict the response value based on predictor variables. Commonly, the structure of ANNs consist of three layers, input layer, hidden layer with several neurons and output layer, with each layer fully connected to the next one through series of weights. The neurons containing transfer functions can receive signals coming from the previous layer and generate outputs which form the input signals for the next layer. The weights can be adjusted based on the empirical data, which make the neural nets having the ability of learning. In the following subsections, three selected types of ANNs namely, BPNN, RBFLN and GRNN, are introduced.

2.2.1. Backpropagation neural network

Multilayer perceptron is the most popular neural networks. As the network is trained using the back propagation method, it is also called BPNN. It is based on searching an optimal set of weights which can generate the minimum error between its outputs and the desired target value. The structure of BPNN consists of 3 types of layers: input layer, hidden layer and output layer. Each layer is fully connected to the next layer. It can be trained by the training data. Each iteration of the training includes two sweeps: feed-forward can generate outputs; backward propagation which calculates the error between the desired target value and the output value to adjust the weights. Moreover it can have more than one hidden layers. But, according to Yesilnacar and Topal (2005), the BPNN with one hidden layer containing enough nodes is capable for most function approximation problem under an acceptable

degree of accuracy.

In an initialized neural network, one training example is involved to run the network in the feed-forward sweep. Firstly, each input node in the input layer transmits the value received forward to each hidden node in the hidden layer. The dot product of all the input nodes and their corresponding interconnection weights performs the input of the corresponding hidden node. Each node in the hidden layer can receive an input from the input layer. Once one hidden node receives the input coming from the input layer, the transfer function in it can yield an output between 0 and +1. The amount of output produced constitutes the new signal that is to be transmitted forward to the output layer. The signals received by the output layer are summarized to generate the neural network solution of fed example, which may have deviation from the target solution caused by the arbitrary selected interconnected weights.

In the backward sweep, the error between the desired target value and the output value can be used to adjust the weights. The feed-forward back propagation process is applied repeatedly till reach the minimum error that the algorithm can produce. But, in some cases, the minimum error is just a local minimum value instead of the global minimum one, which is the drawback of the BPNN. Moreover, as stated by Dreiseitl and Ohno-Machado (2002) the BPNN also has the problem of overfitting, which means that the algorithm memorize the data set instead of finding the underlying distribution. However, it can be solved by using a method called weight decay, which shrink the regression coefficients by adding a penalty to the error function (Dreiseitl and Ohno-Machado 2002; Hastie et al. 2005).

The structure of BPNN, which has three neurons in the input layer and three neurons in the hidden layer, is shown in figure 2-1. The a_i represents the weight that connects the neuron in the hidden layer and the output. Additionally, the circles with label of "+1", which correspond to the intercept term, are called bias units.

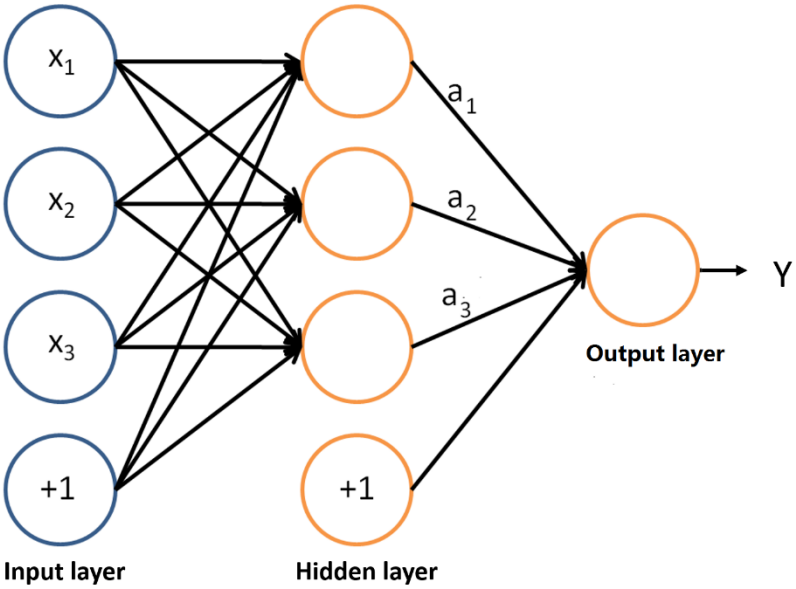


Figure 2-1 the structure of backpropagation neural network.

2.2.2. Radical basis functional link nets

RBFLN can be used for nonlinear input-output relationships (Looney 2002). An RBFLN consists of three layers: (I) an input layer with n nodes, which receives values from observation data; (ii) a hidden layer which contains m artificial neurons; and (iii) an output layer (Looney 2002). The active function in the neurons in hidden layer is a radial basis function (RBF) (Looney 2002) and the hidden layer and the output layer are connected through a series of weights. The equation 3 shows a usual Gaussian RBF where the x presents an n-dimensional vector, v present a center, and σ is the spread parameter(Looney 2002).

$$Y = f(x; v) = \exp\{- ||x - v||^2/2*\sigma^2)\} \qquad \text{Equation 3}$$

The structure of RBFLN, which has n neurons in the input layer and m neurons in the hidden layer, is shown in figure 2-2. In an initialized RBFLN, one training data received by the input layer is linked to all the neurons in the hidden layer. The data transmitted to the hidden layer is fed into the radial basis function to generate an output. Then the output generated by the hidden layer and the weight-A (A_1, A_2..., and A_m) form a dot product, which is transmitted to the output layer. Additionally, the input data is also linked to the output layer through an additive linear model with

a set of weight-B (B_1, B_2..., and B_n). Finally, the output layer responses the output of fed data. The desired weight A and weight B of the RBFLN can be trained using gradient descent approach (Looney 2002).

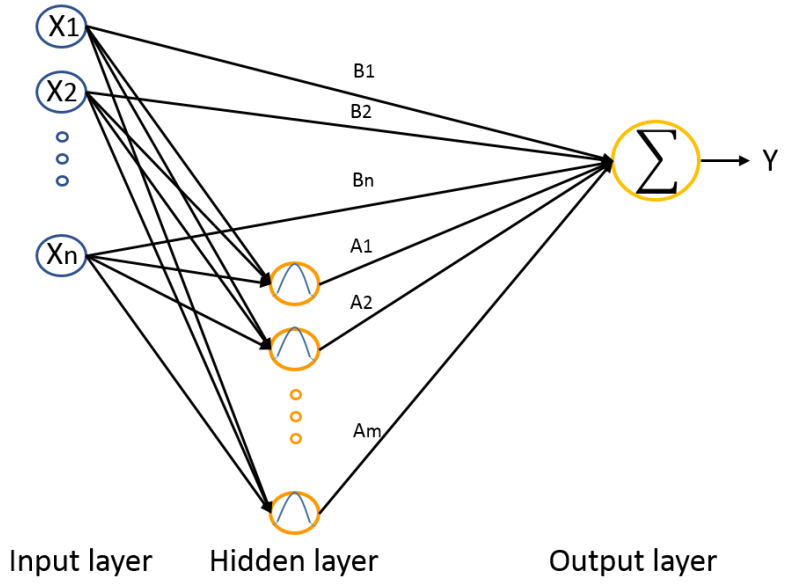


Figure 2-2 RBFLN with n neurons in the input layer and m neurons in the hidden layer.

2.2.3. General regression neural network

GRNN is a memory-based supervised feed-forward network, which is based on nonlinear regression and designed for function approximation (Fung et al. 2005). The algorithm has a highly parallel structure, and even with sparse data in a multiple dimension space, the algorithm can provide smooth transitions from one input data to another (Specht 1991).

GRNN is a 3-layer network that has an input layer, a hidden layer consisting of pattern units and summation units, and an output layer. The p-dimensional input data consists the p input units which are fed to the pattern units in hidden layer. Then those inputs are calculated through the transfer function in each pattern units. Moreover, the Gaussian kernel can be used as the transfer function (Specht 1991). Then, the outputs of all the pattern units are connected to two summation unites through two sets of weights, A_i and B_i. Finally, the output just divides summation unit A by summation unit B to generate the estimate of Y. The figure 2-3 shows the structure

of GRNN. The input is an n-dimensional vector and the number of units in the pattern units is m.

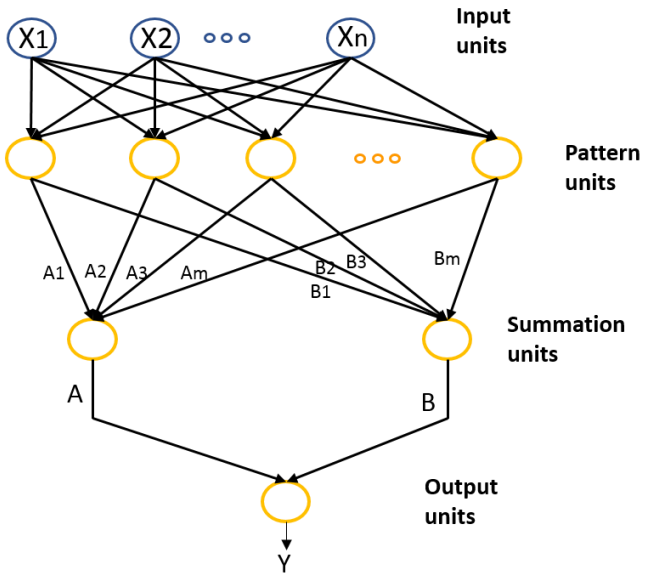


Figure 2-3 the structure of GRNN.

The transfer function, which is the Gaussian kernel, in each pattern unite contains two parameters, one is center C_i and the other one is smooth parameter σ. The Gaussian kernel is shown in equation 4.

$$K(||x_i - C_i||) = \exp\{-||x_i - C_i||^2/2*\sigma^2)\} \quad \text{Equation 4}$$

The centers C_i and the weights A_i and B_i can be generated from the training data (Specht 1991). Therefore, given a training data set and an independent data set, the transfer function is optimized by the selection of a single σ for all pattern unites, which is the common spherical or radial basis function kernel band width. In most situations, the GRNN has a unique σ that can generate the minimum Mean Square Error (MSE) between the predicted output and the desired output (Specht 1991).

2.3. Background of visceral leishmaniosis

Visceral leishmaniosis, which is also known as kala-azar, black fever, is a zoonotic, vector-borne, infectious disease transmitted to humans through the bite of infected

sand flies (Tsegaw et al. 2013). It is a disease caused by protozoan parasites of the Leishmanial genus. This disease is the second-largest parasitic killer in the world (after malaria). It is the ninth most infectious disease which infects about 200,000 to 400,000 people each year worldwide (WHO 2016b).

Numerous researches have shown that environmental, demographic, and statistical data of the ecology of VL can be integrated to make susceptibility map for VL (Peterson and Shaw 2003; Castillo-Riquelme et al. 2008; Salahi-Moghaddam et al. 2010). Moreover, Rajabi et al. (2014) have proved that it is feasible to use the 7 kinds of evidence maps which are temperature, proximity to river, precipitation, proximity to nomadic villages, land cover, altitude, proximity to health-centers to predict the probability of the presence of VL in north-western Iran.

2.4. Using of LR and ANNs in GIS

A number of studies have been carried out using LR and ANNs in GIS, while most of them focused on landslide study.

Lee and Sambath (2006) used logistic regression to link factors, such as slope, curvature, and distance from drainage, etc., with the occurrence of landslide to make landslide susceptibility map;

Ayalew and Yamagishi (2005) use logistic regression to predict the presence and absence of landslide occurrence. And according to the result of the regression, five categories of landslide susceptibility were made.

Yilmaz (2009) linked landslide-related factors, such as geology, faults, slop and angle etc., to landslide using LR, BPNN and another method called frequency ratio. It shown that all of those methods can generate a good result, meanwhile, the BPNN had the best performance.

Dai and Lee (2002) used logistic multiple regression to model the relationship between landslide and pertinent landslide characteristics. They stated that the logistic regression has advantage in predict the occurring or absent of an event, and the

predictive value can be interpreted as probability.

Porwal et al. (2003) applied RBFLN to base metal deposit potential mapping in the Aravalli province and showed that the spatial distribution of the high favorability zones predicted by the RBFLN model is consistent with the conceptual models of base metal metallogeny in the study area.

Kumar and Anbalagan (2015) generated landslide inventory maps in Tehri reservoir rim area using RNFLN. The landslide causative factors were derived from remote sensing and the accuracy of the prediction was found to be 86%.

Yesilnacar and Topal (2005) applied logistic regression and back-propagation neural networks to produce landslide susceptibility map in a medium scale and showed that the susceptibility map produced by back-propagation is more realistic. Moreover, the logistic regression was proved to be helpful in eliminating unrelated inputs factors and finding out the significance of the related significant ones.

Rajabi et al. (2014) developed an environmental modelling of VL by susceptibility-mapping using RBFLN. The study showed that the RBFLN model can generate reliable susceptibility maps without any external knowledge.

Singer and Kouda (1996) used a feedforward neural network with one hidden layer and five neurons to identify the distance to Kuroko mineral deposits. And, the method succeeded in recognize all of the known deposits which are expected to be identified.

Brown et al. (2000) applied BPNN to estimate the favorability for gold deposits using a raster GIS database and showed that the BPNN has several advantages over conceptual fuzzy-logic method and empirical weight of evidence.

Beucher et al. (2013) used RBFLN for soil mapping in the area of Sirppujoki River catchment and showed that this method has good predictive classification abilities regarding mapping acid sulfate.

Fung et al. (2005) compared four types of neural networks, including GRNN, BPNN, PNN and Probabilistic Neural Network (PrNN) for the prediction of mineral prospectively and showed that the PNN, GRNN and PrNN all have a better performance than BPNN.

3. Case study of visceral leishmaniosis

3.1. Data resource of visceral leishmaniosis

In the case study, a GIS data set, which was generated by Rajabi et al. (2014), in the area of Kalaybar and Ahar, Iran, was used to obtain the social-environmental feature vectors associated with VL. It consists of 7 maps in raster format (Fig. 3.1). All maps, which already have the same scale of 1:50,000, were preprocessed using ArcGIS 10.2 to make them covering the same area.

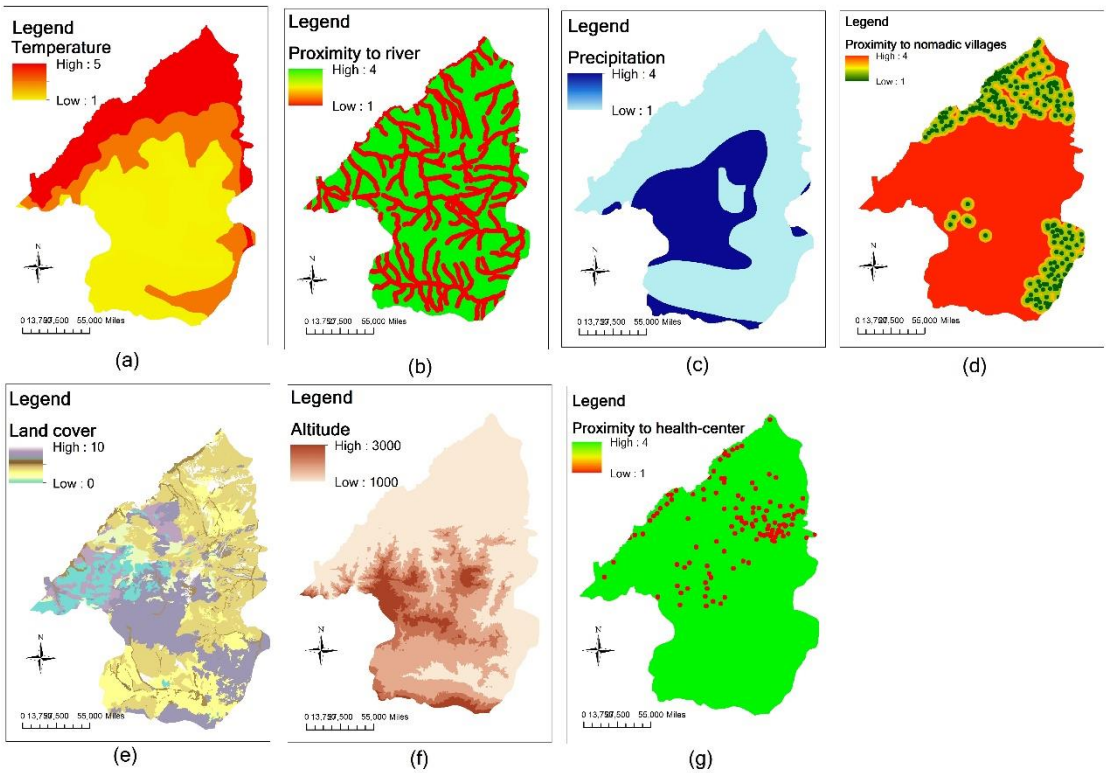


Figure 3-1.Evidence map for VL. (a) Temperature, (b) Proximity to river, (c) precipitation, (d) Proximity to nomadic villages, (e) Land cover, (f) Altitude, (f) Proximity to health-centers.

In addition, one point data set (Fig. 3.2) containing the presence/absence of VL in 59 villages of the study area, were used to generate the training and validation datasets. The dataset contains 30 endemic points that show the presence of VL, and 29 non-endemic points exhibiting the absence of the VL. Ten out of the 59 points (5 points from each) were selected as validation data. Each point contains an attribute value 0 or 1 where 0 means low susceptibility and 1 means high susceptibility.

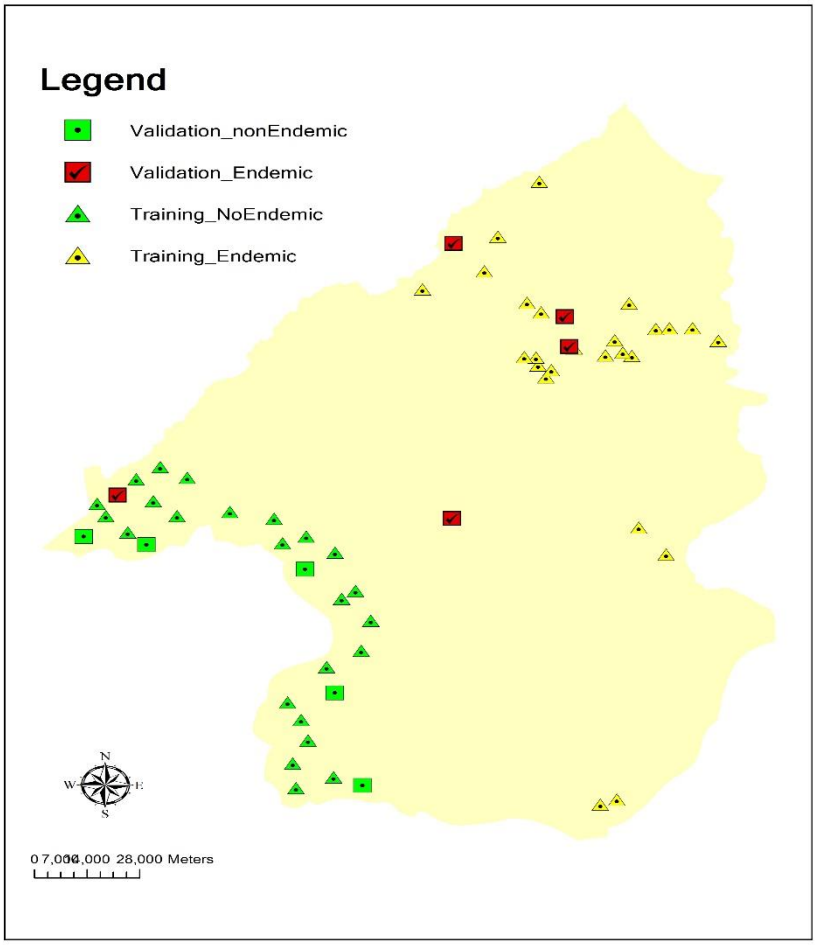


Figure 3-2.The training and validation points.

3.2. Implementing the models on visceral leishmaniosis

The models were developed using Python. The figure 3.3 shows the general process of producing VL susceptibility map and the validation of the models.

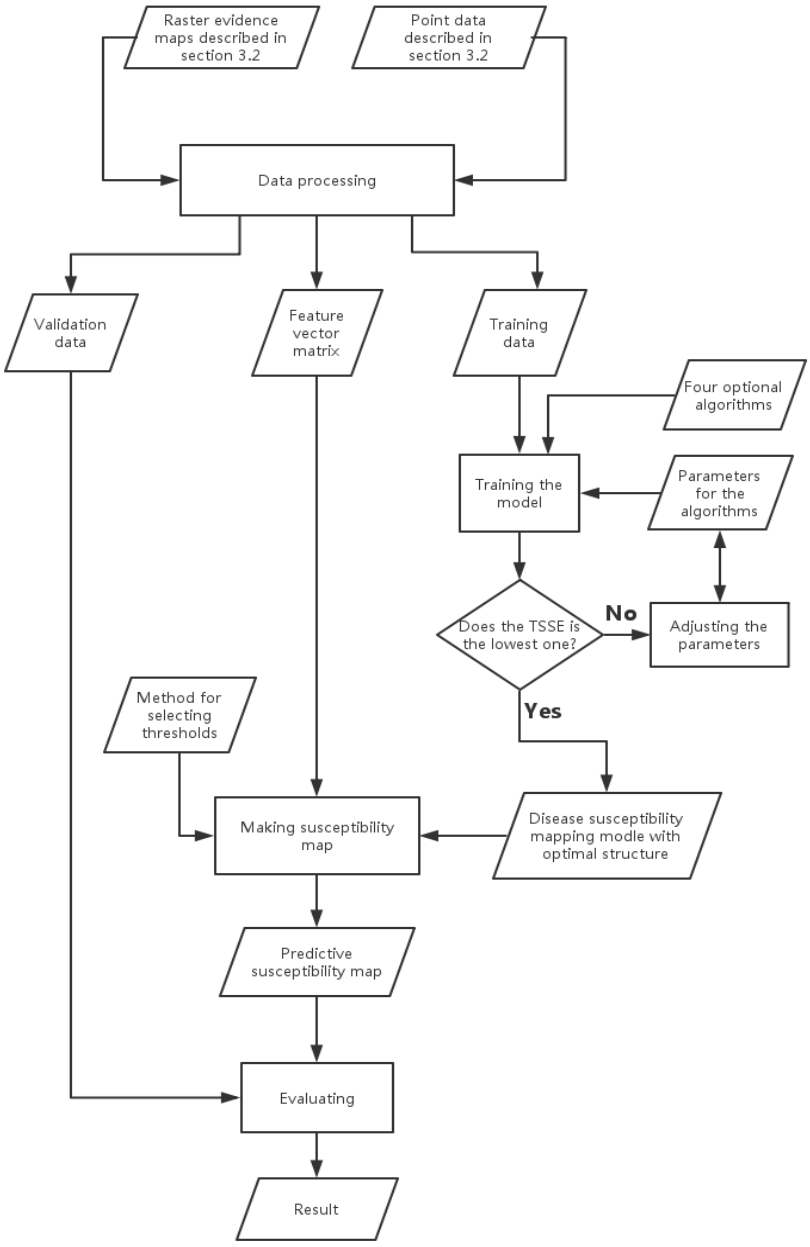


Figure 3-3.The general process of producing VL susceptibility map and the validation of the models.

The process of making disease susceptibility map using these models is described below.

The raster dataset containing the social environmental information and the point dataset, which stores the coordinates of the training and validation points and the attribute (0 or 1) of each point, were connected to the model. All the raster maps were integrated together to form a feature vector matrix to be used as the input feature vector for the model. Each element in the feature vector matrix contained a 7-dimensional vector consisting of the corresponding value derived from the seven evidence maps. The column and row number of the elements in the matrix could be used to locate its associated position in the raster maps. All the values of the raster maps were normalized to [0, 1] using $X_n = (X - X_{min}) / (X_{max} - X_{min})$ in which X_n is the normalized value and X_{min} and X_{max} are the minimum and maximum value of the corresponding map respectively. In this way, the land cover, which is nominal data, was converted to values in the range of 0 and 1. The value presents the classification of the land cover.

The social-environmental feature vector of each training and validation point was then derived from the feature matrix according to corresponding coordinates stored in the point shapefile. The derived feature vector and the corresponding attribute value of one point formed a pair of input and desired output. All pairs formed the training and validation dataset. The model needed to be trained under different parameters which influenced the model's performance. The parameters that produced the lowest training sum of squared error (TSSE) were to be chosen in generating the predictive susceptibility value of all the cells in the study area. Four algorithms (LR, BPNN, RBFLNN and GRNN) were implemented for predicting the susceptibility value.

A set of thresholds were selected to classify the cells into 5 categories, which are very low susceptibility area, the low susceptibility area, the moderate susceptibility area, the high susceptibility area and the very high susceptibility area. The thresholds were selected by using reclassify function in ArcGIS 10.3. The reclassify method selected was natural break.

3.3. Validation of the models

After the susceptibility maps were produced, two aspects of model performance were tested. The first was discrimination, which tested how good the endemic and non-endemic area can be divided; the second was calibration, which tested how accurate the validation points are classified.

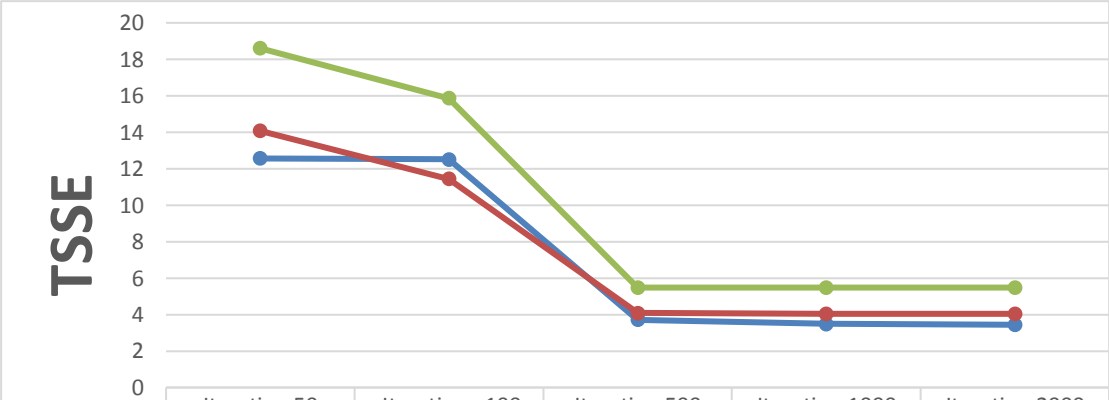
The receive operating characteristic (ROC) curve was used to measure the discrimination ability of these models (Dreiseitl and Ohno-Machado 2002). It is a plot of true positive rate on the y-axis versus false positive rate on the x-axis. The false positive rate is defined by FP / (FP + TN) and the true positive rate is calculated by TP/ (TP + FN) where FP = false positive, FN = false negative, TP = true positive and TN = true negative. The area under the curve (AUC) can indicate the quality of the model regarding the aspect of discrimination. The AUC is between 0 and 1 (1 means that the results are all correct and 0 means the results are totally incorrect), and was generated by cross-validation. All the data including training and validation were divided into six groups. Each time, one group of data was taken out as the test data to get the predictive value. The process was repeated until all has been tested. The ROC curve of all the predictive value was then generated.

The calibration was measured by plotting the validation points onto the curve of predictive value versus cumulative percentage of area.

4. Result

4.1. Result of logistic regression

Two parameters influences the performance of LR model: learning rate (η) and the number of iterations. Thus, LR models with different learning rate and iteration were explored. The model with the learning rate of 0.05 and 2000 iterations got the lowest TSSE. The TSSEs generated by the models with different learning rates and iterations are shown in Fig. 4.1.



	Iteration 50	Iteratione 100	Iteration 500	Iteration 1000	Iteration 2000
η 0.05	12.56	12.51	3.71	3.48	3.44
η 0.1	14.07	11.44	4.08	4.04	4.04
η 0.2	18.6	15.87	5.48	5.48	5.48

Figure 4-1.The training sum of squared error (TSSE) versus number of iterations for different learning rates

As shown in Fig. 4.2, the study area was divided into 5 zones: very low susceptibility area, the low susceptibility area, the moderate susceptibility area, the high susceptibility area and the very high susceptibility area. This was conducted by selecting a series of thresholds which were 0.12, 0.27, 0.47 and 0.71. The very low and low susceptibility area, which covered 50.5% of the study area, contained 91.6% of the non-endemic training points and 8% of endemic training points. 84% of the endemic and 0% of the non-endemic training points were categorized into the high and very high susceptibility area respectively, which comprised 34.25% of the study

area. The moderate susceptibility class contained 8% of the endemic training points and 8.4% of the non-endemic training points. The Fig. 4.4 and 4.3 show the calibration and discrimination abilities of this model.

As shown in Fig. 4.3, 80% of the non-endemic validation points were classed into very low susceptibility area. The rest was in the low susceptibility area. Twenty percent of endemic validation points were located in the low susceptibility area. Also, 60% of the endemic validation points were categorized into high and very high susceptibility area, and 20% of the endemic susceptibility points were in the moderate susceptibility area. The same information is also shown in Fig. 4.5. The ROC curve for this model, which produced the AUC of 0.899, is shown in Fig. 4.4.

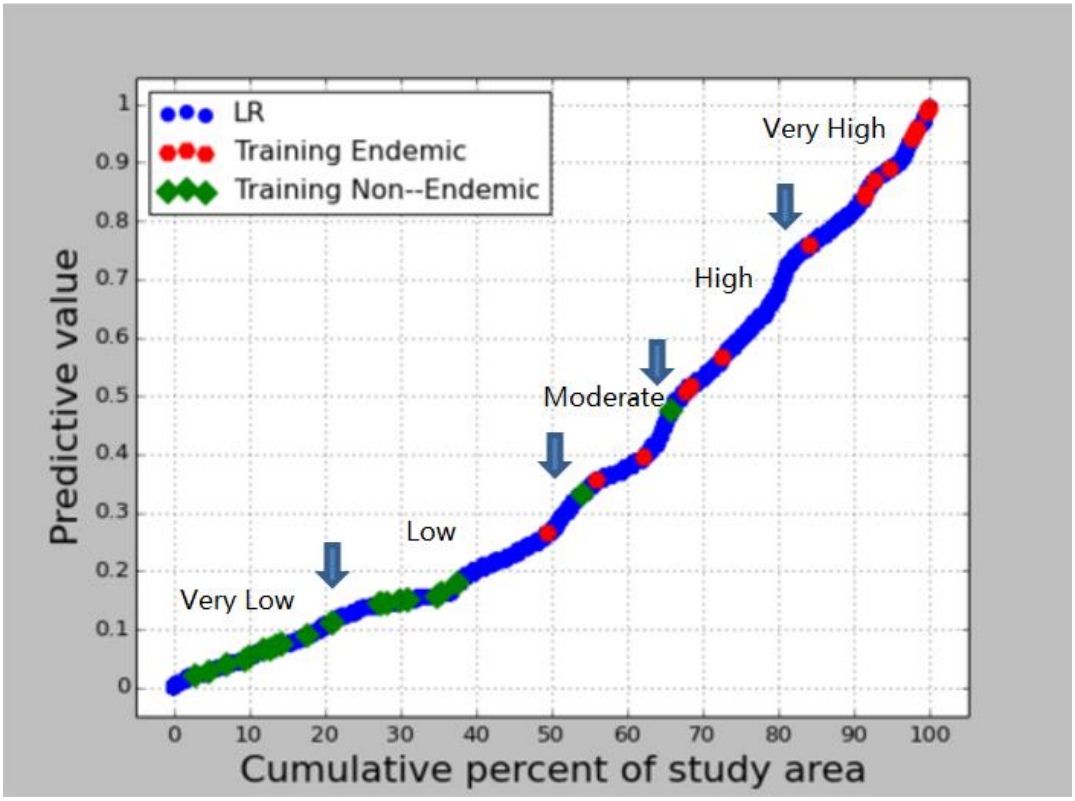


Figure 4-2.Predictive susceptibility of training data versus cumulative percent of the study area for LR. The blue arrows mean the locations of the thresholds.

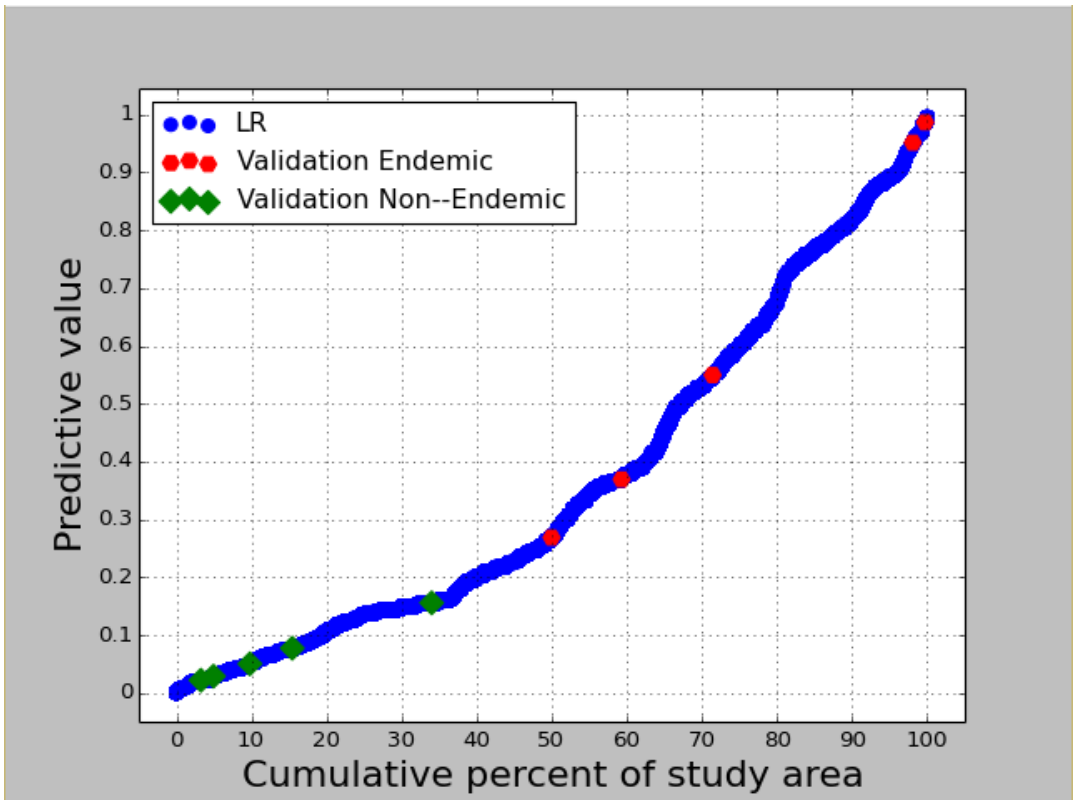


Figure 4-3.Predictive susceptibility of validation points versus cumulative percent of the study area for LR.

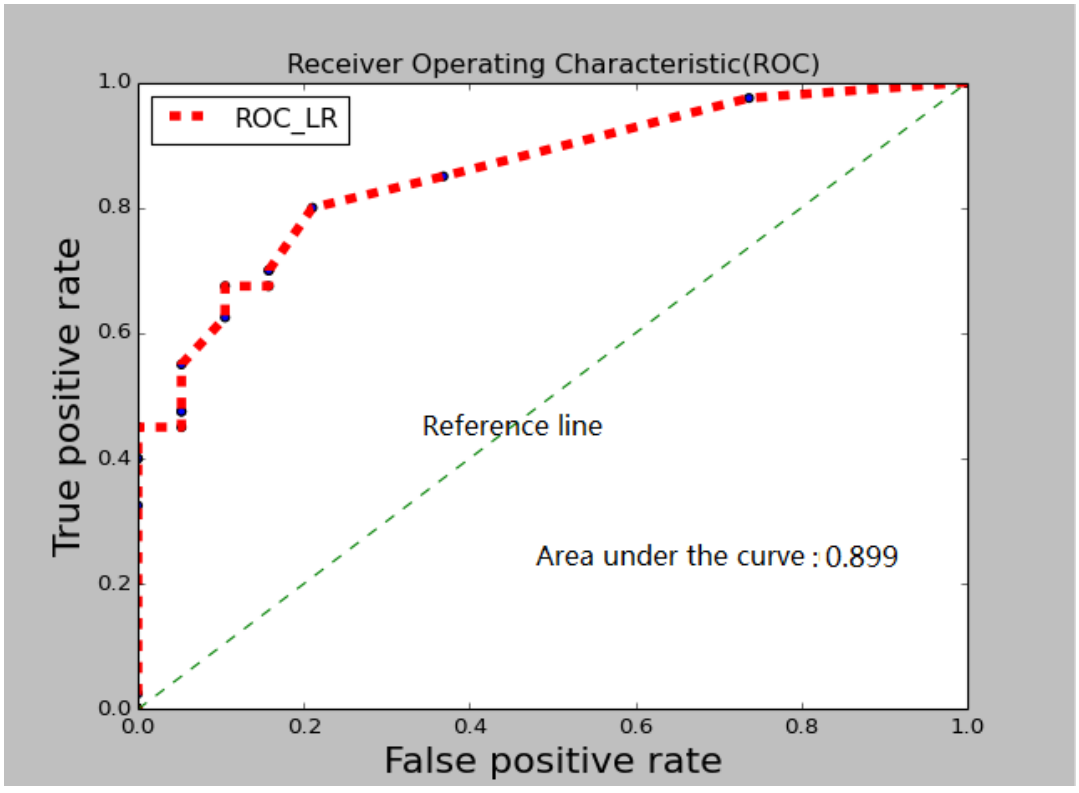


Figure 4-4. ROC for validation of LR model.

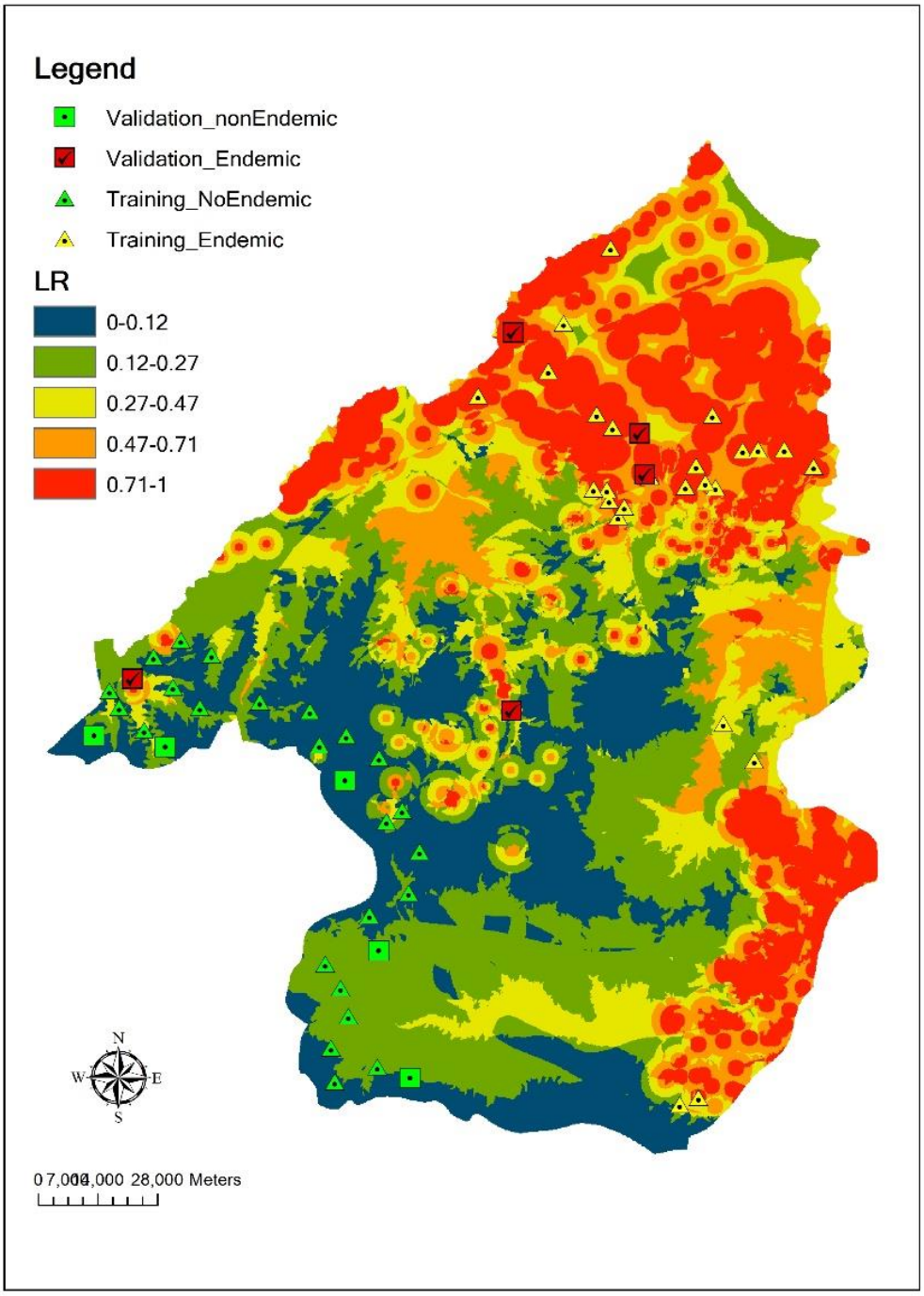
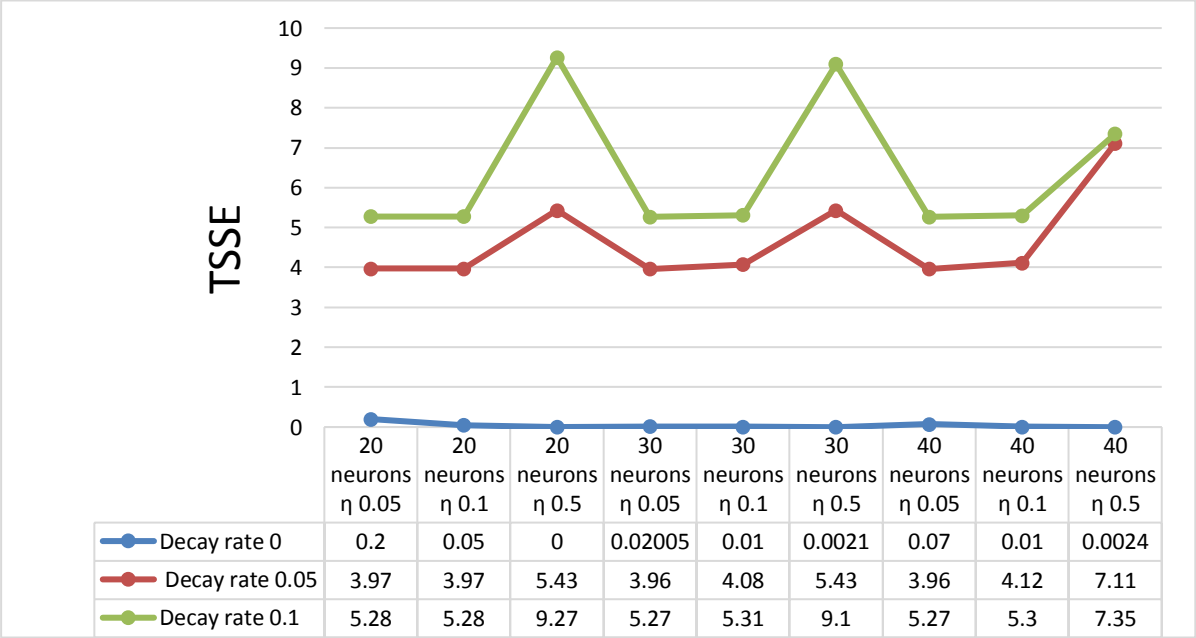


Figure 4-5.VL susceptibility map produce by LR.

4.2. Result of backpropagation neural network

The performance of BPNN is affected by the number of nodes in the hidden layer, the learning rate, and the weight decay. The TSSEs of the BPNN models with different parameters are shown below. All the training had the maximum training iteration of 10,000, and were processed five times to get the minimum TSSE. The model with 20 neurons, the learning rate (η) of 0.5 and the weight decay of 0.0 was selected. While considering the overfitting problem of BPNN, another model with 30 neurons in the hidden layer, the learning rate of 0.05 and the weight decay of 0.05, was chosen to test if the model with a weight decay element can have a better performance. The TSSEs generated by the BPNN models with different structures are shown in Fig. 4.6.



	20 neurons η 0.05	20 neurons η 0.1	20 neurons η 0.5	30 neurons η 0.05	30 neurons η 0.1	30 neurons η 0.5	40 neurons η 0.05	40 neurons η 0.1	40 neurons η 0.5
Decay rate 0	0.2	0.05	0	0.02005	0.01	0.0021	0.07	0.01	0.0024
Decay rate 0.05	3.97	3.97	5.43	3.96	4.08	5.43	3.96	4.12	7.11
Decay rate 0.1	5.28	5.28	9.27	5.27	5.31	9.1	5.27	5.3	7.35

Figure 4-6.TSSE generated by BPNN model with different parameters.

Table 4.2 and 4.3 show the predictive values of the validation points and the test sum of squared errors generated by the two BPNNs respectively. The network with the weight decay of 0.05 generated a lower test sum of squared error. Moreover, the network without the weight decay could not totally separate the validation points, while the network with the weight decay could. The same information is also shown in Fig. 4.7 and Fig. 4.9. Therefore, the model with 30 neurons in the hidden layer, learning rate of 0.05, and weight decay of 0.05, was selected for further analysis.

Table 4.1.Predictive values of the validation points generated by the two BPNNs (predictive value 1 means the BPNN model with weight decay and the predictive value 2 represents the one without weight decay)

N.	State	Class	Predictive value 1	Predictive value 2
1	Endemic	1	0.3249	0.7165
2	Endemic	1	0.4993	1.0146*10-3
3	Endemic	1	0.5947	9.9999
4	Endemic	1	0.9239	9.9999
5	Endemic	1	0.9752	1
6	Non-Endemic	0	0.095	1.0146*10-6
7	Non-Endemic	0	0.2879	5.9996*10-3
8	Non-Endemic	0	0.1612	1.2014*10-5
9	Non-Endemic	0	0.1333	2.223*10-6
10	Non-Endemic	0	0.1734	1.1043*10-5

Table 4.2.Test sum of squared errors generated by the two selected BPNN models

BPNN	test sum of squared error
Weight Decay=0.05	1.04
No Weight Decay	1.07

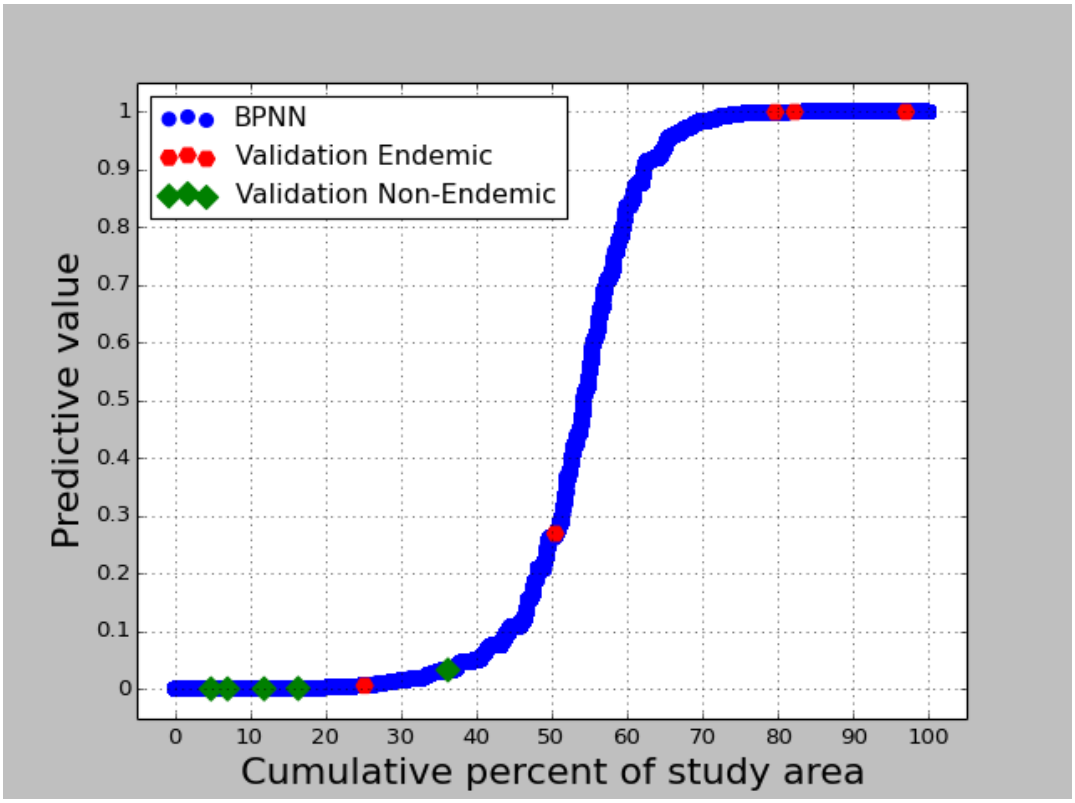


Figure 4-7.Predictive susceptibility of validation points versus cumulative percent of the study area for BPNN model with 20 neurons, the η of 0.5 and the weight decay of 0.0.

As is shown in Fig. 4.8, the study area was divided into 5 zones by selecting a series of threshold namely 0.21, 0.38, 0.57 and 0.76. The five categories included very low susceptibility area; low susceptibility area; moderate susceptibility area; high susceptibility area; very high susceptibility area. The very low and low susceptibility area, which covered 52.5% of the study area, contained 91.7% non-endemic training points and 8% endemic training points. 72% and 0% endemic and non-endemic training points were categorized in high and very high susceptibility area respectively,

which comprised 28.2% of the study area. The moderate susceptibility class contained 20% endemic training points and 8.3% non-endemic training points.

Eighty percent of the non-endemic validation points were classed into the very low susceptibility area, and the rest were in the low susceptibility area (Fig. 4.9). There were 20% endemic validation points located in the low susceptibility area. Also, 60% of the endemic validation points were categorized into high and very high susceptibility area and 20% of the endemic susceptibility points were in the moderate susceptibility area. The same information is also shown in Fig. 4.11. The ROC curve for this model, which produced the AUC of 0.942, is shown in Fig. 4.10.

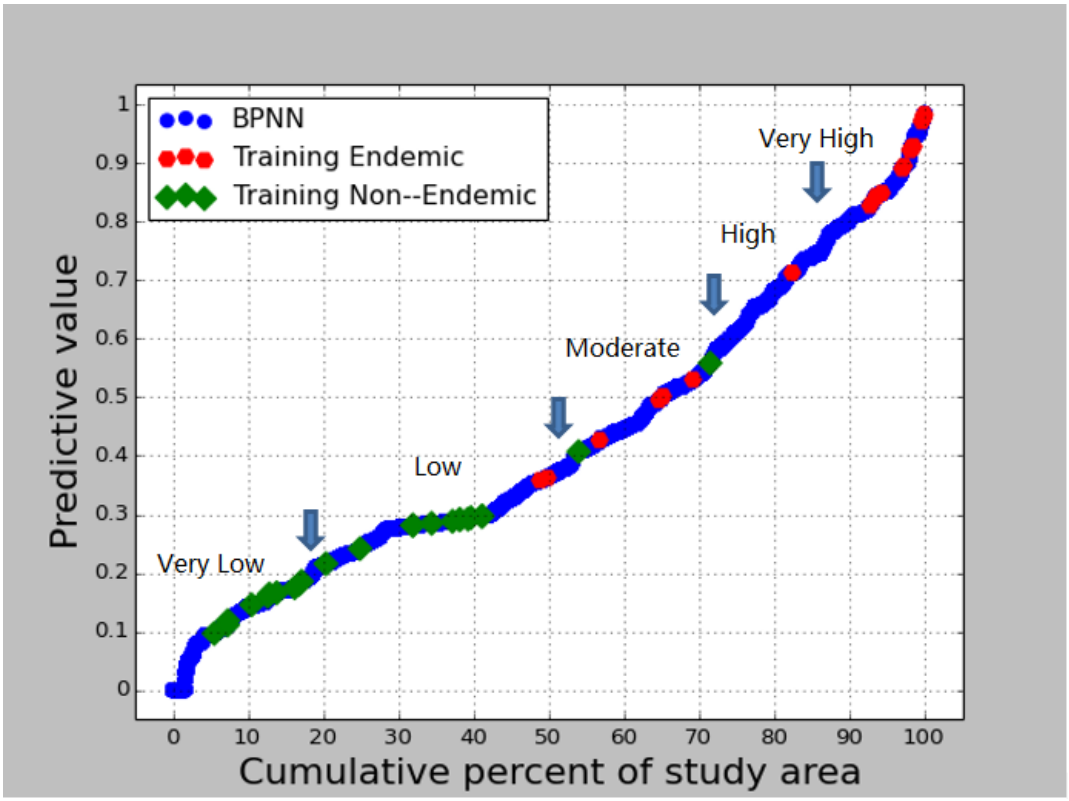


Figure 4-8.Predictive susceptibility of training points versus cumulative percent of the study area for BPNN model with 30 neurons, learning rate of 0.05, and weight decay of 0.05.

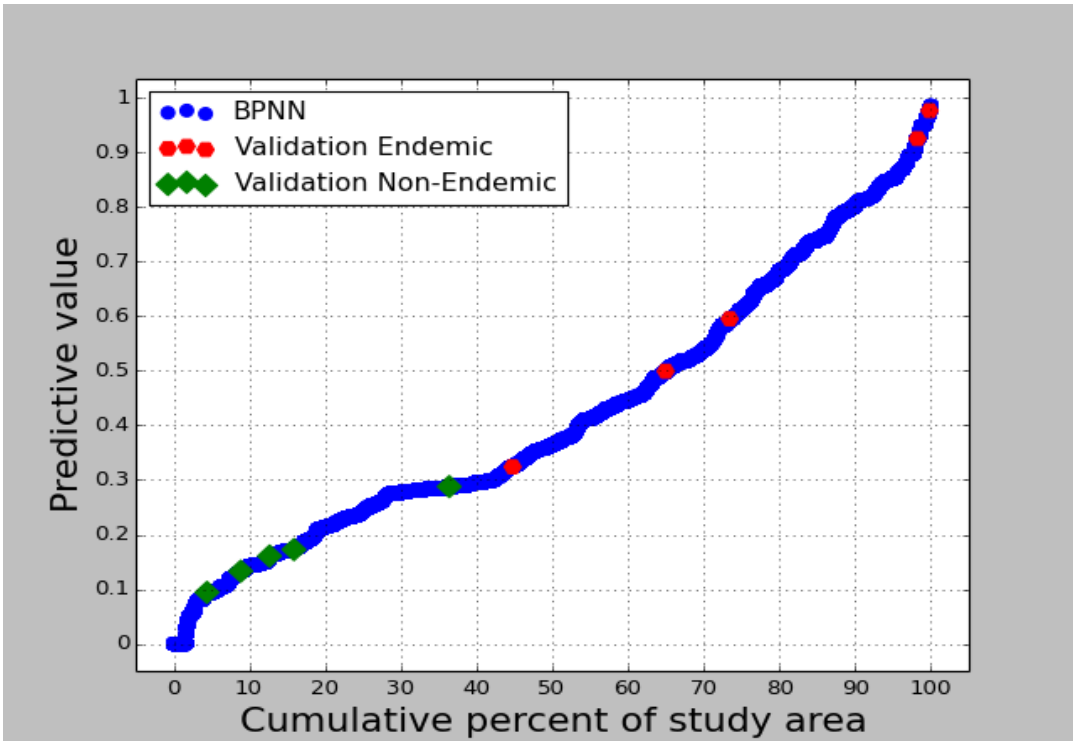


Figure 4-9.Predictive susceptibility of validation points versus cumulative percent of the study area for BPNN model with 30 neurons, learning rate of 0.05, and weight decay of 0.05..

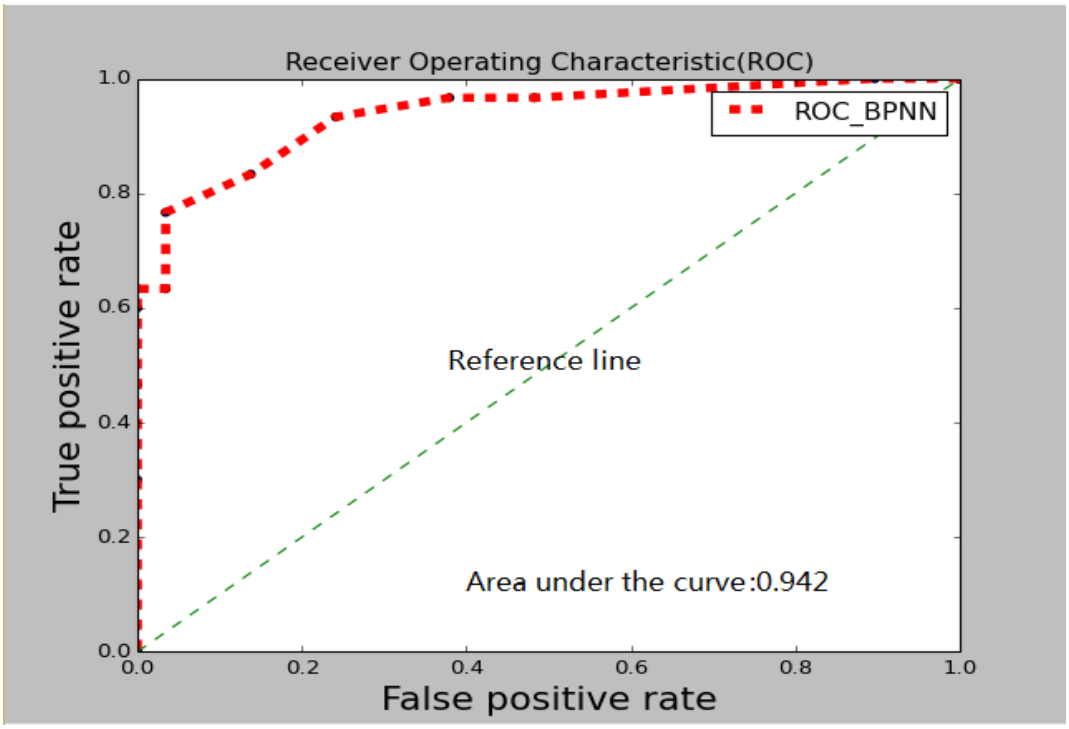


Figure 4-10. ROC for validation of BPNN model.

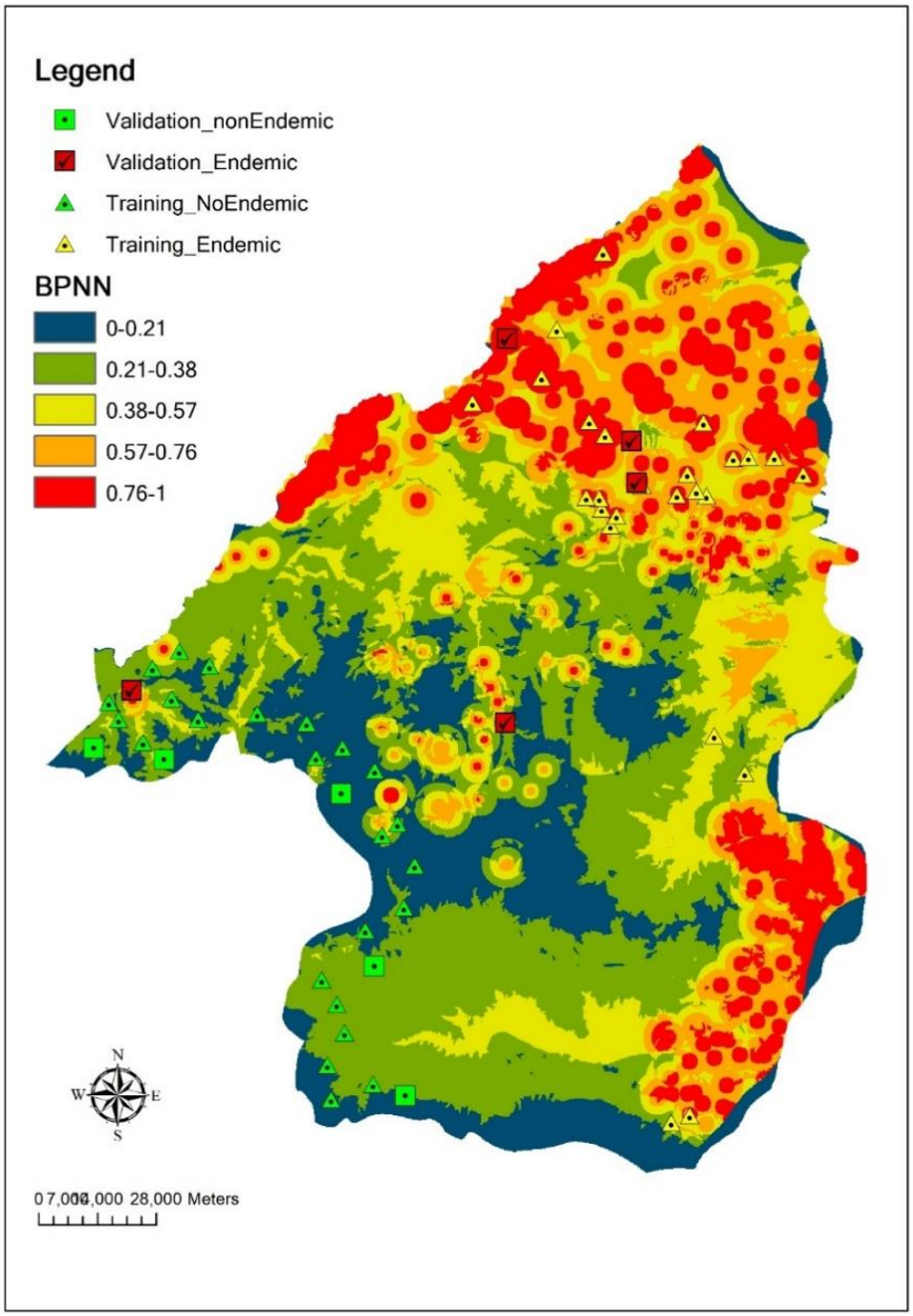


Figure 4-11.VL susceptibility map produce by BPNN.

4.3. Result of radical basis functional link nets

The structure of RBFLN is only determined by the number of neurons in the hidden

layer as enough iterations were reached. The TSSEs produced by the RBFLN models with different structures are shown in Fig. 4.12. The RBFLN model with 30 neurons generated the lowest TSSE of 2.96 at the training iteration of 2000.

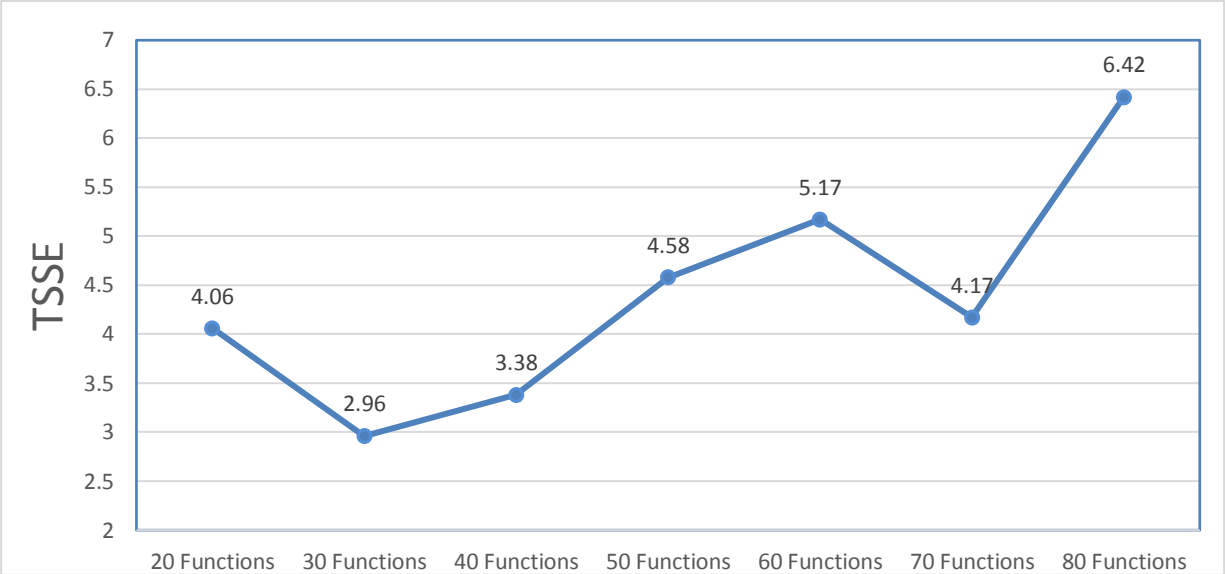


Figure 4-12.TSSEs produced by the RBFLN models with different structures.

As shown in Fig. 4.13, the study area was divided into five zones by selecting threshold values of 0.29, 0.402, 0.552 and 0.719, generating five categories (very low susceptibility area, low susceptibility area, moderate susceptibility area, high susceptibility area and very high susceptibility area). The very low and low susceptibility area contained 91.7% of the non-endemic training points and none of the endemic training points. Eighty-eight percent of endemic training points were in the high and very high susceptibility area. The moderate susceptibility class comprised 12% endemic training points and 8.3% non-endemic training points.

The Fig. 4.14 shows that 100% of the non-endemic validation points were classed into very low and low susceptibility area, covering 44.2% of the study area. There were 20% endemic validation points located in the low susceptibility area. Also, 60% of the endemic validation points were categorized into high and very high susceptibility area, which comprised 35.8% of the study area. Forty percent of the endemic susceptibility points were in the moderate susceptibility area. The same information is also shown in Fig.4.16. The ROC curve for this model, which produced the AUC of 0.938, is shown in Fig. 4.15.

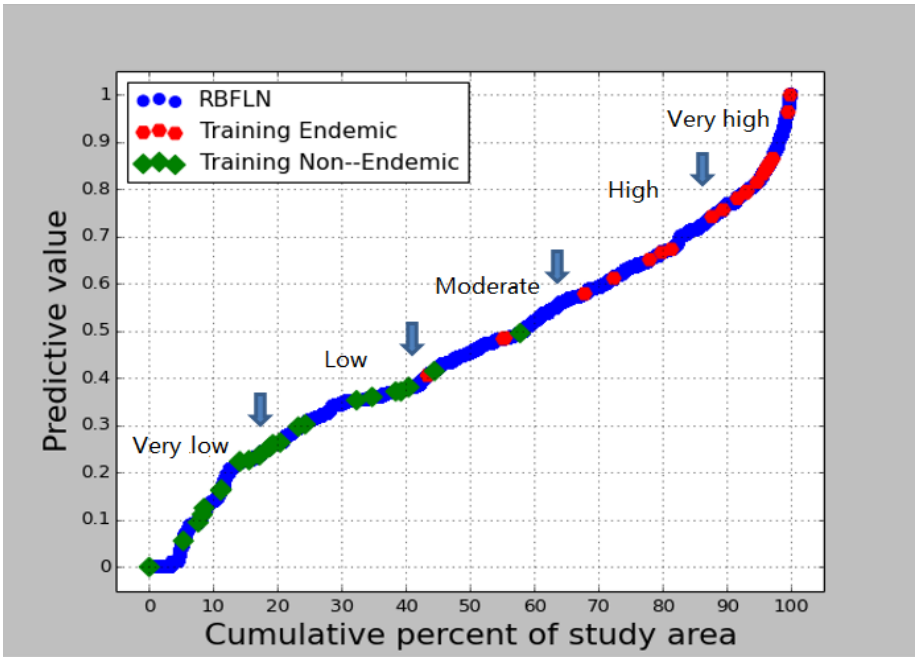


Figure 4-13.Predictive susceptibility of training points versus cumulative percent of the study area for RBFLN model.

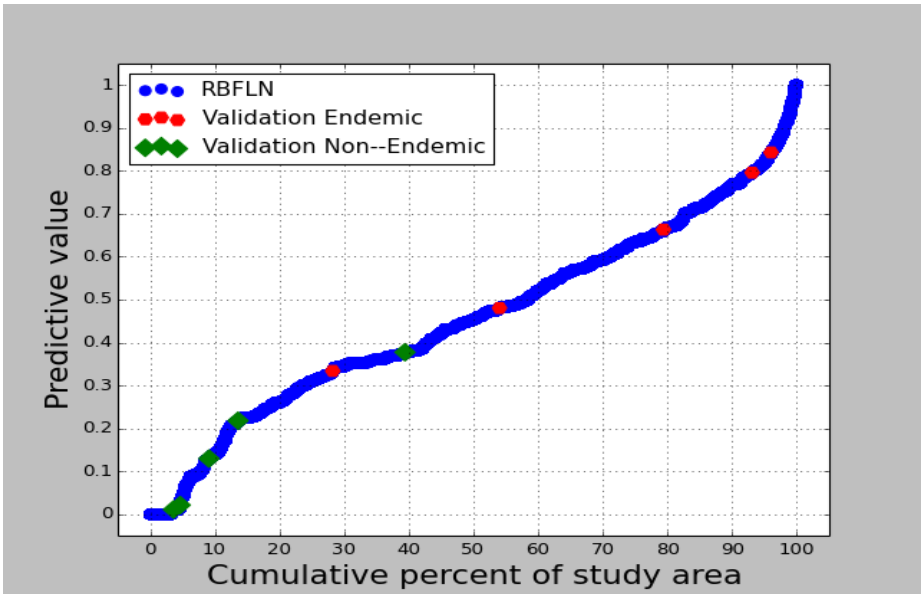


Figure 4-14.Predictive susceptibility of validation points versus cumulative percent of the study area for RBFLN model.

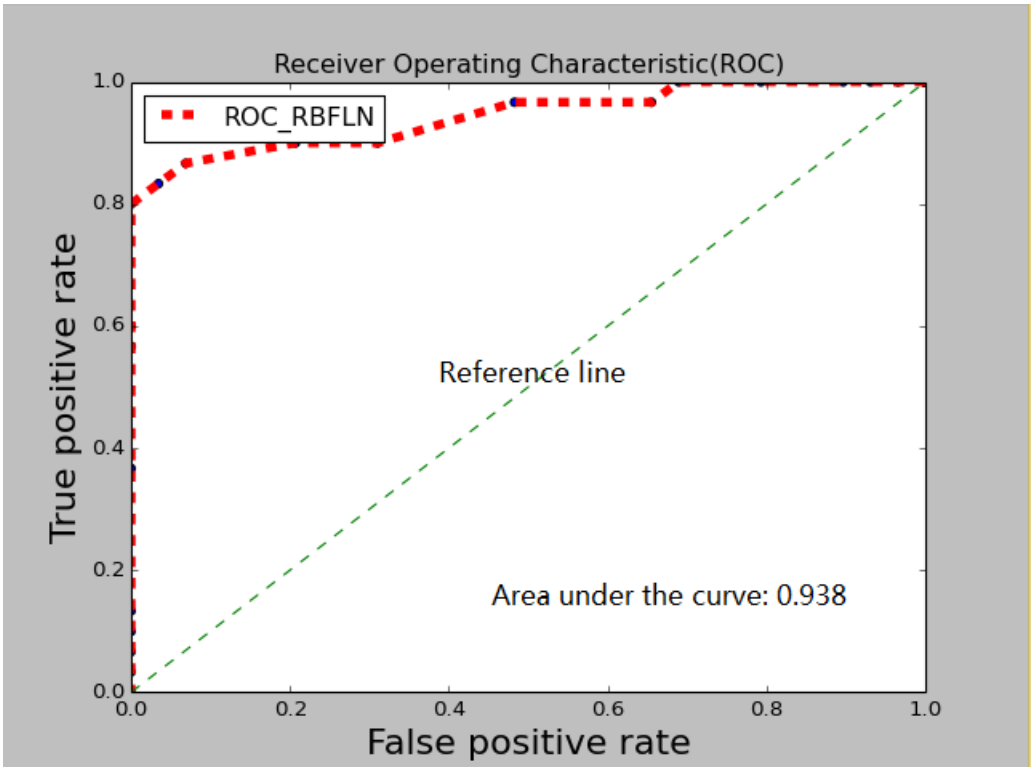


Figure 4-15.ROC for validation of RBFLN model.

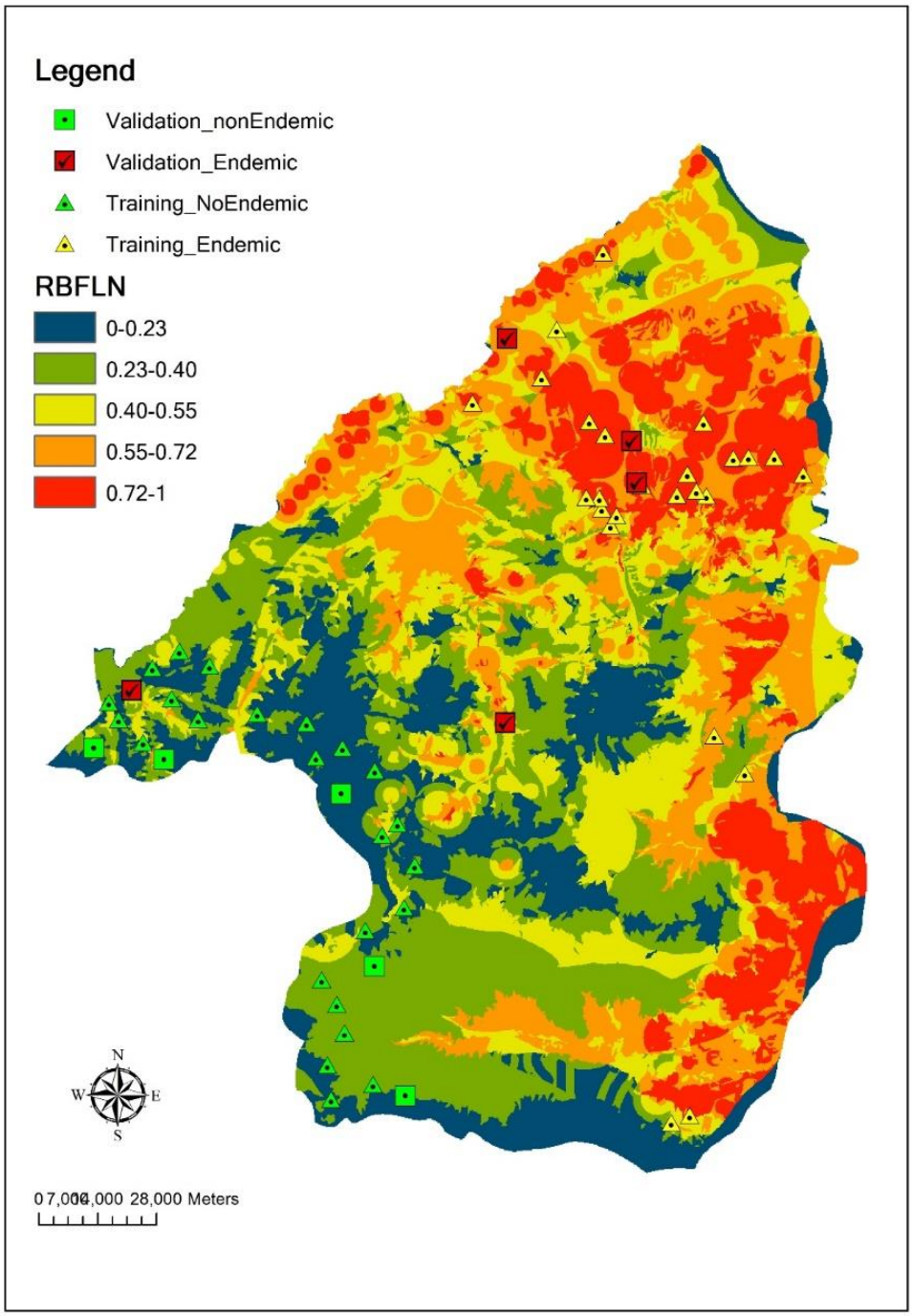


Figure 4-16.VL susceptibility map produce by RBFLN.

4.4. Result of general regression neural network

The performance of GRNN is effected by the spread constant (σ). But, the training procedure for GRNN model is different from the other three algorithms. It can be described as follows: firstly, whole training sample was divided into two parts, the training sample and the test sample. GRNN was then applied on the test data based on the training data, and the sum of squared error (SSE) for different σ was calculated. Finally, the minimum SSE and corresponding value of σ were determined. In this case study, the training data were divided into 5 groups and in each training iteration, one group of data was selected as the test data so that 5 optimal σ were found. The optimal σ was obtained by calculating the average of those 5 σ .

Fig. 4.17 shows the SSE versus σ in the 5 iterations. The minimum SSE and corresponding value of σ are shown in Fig. 4.18.

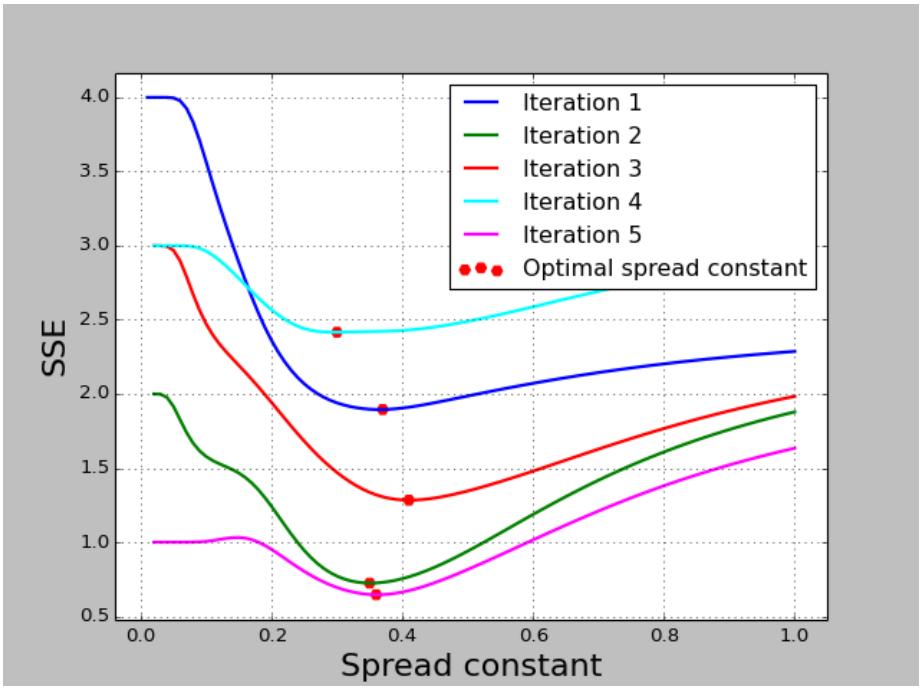
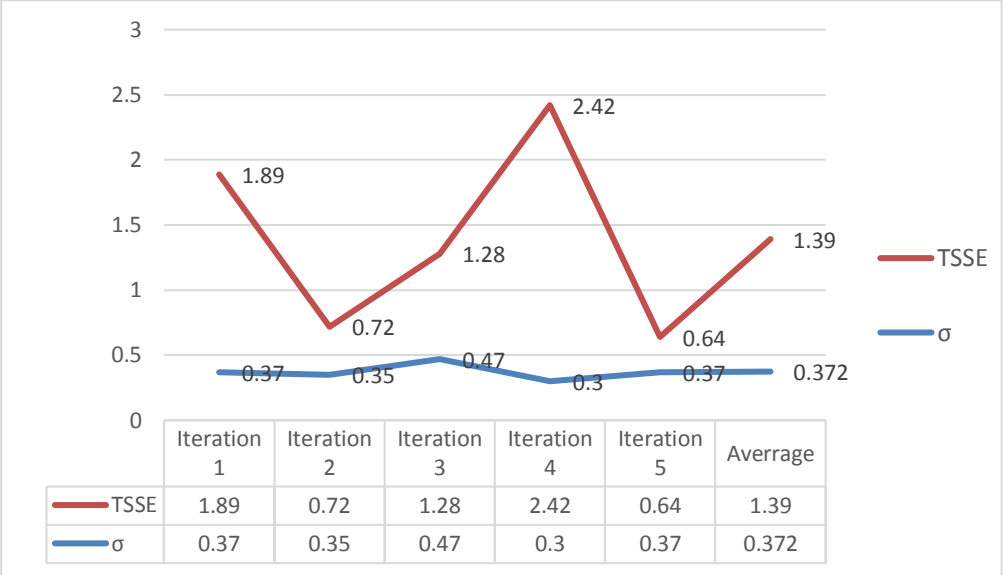


Figure 4-17.The SSE versus σ in the 5 iterations



	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Averrage
TSSE	1.89	0.72	1.28	2.42	0.64	1.39
σ	0.37	0.35	0.47	0.3	0.37	0.372

Figure 4-18.The TSSE and σ in the 5 iterations.

As shown in Fig. 4.19, the threshold values of 0.21, 0.35, 0.53 and 0.77, were used to classify the values into five categories (very low susceptibility area, low susceptibility area, moderate susceptibility area, high susceptibility area and very high susceptibility area). The very low and low susceptibility area contained 95.8% non-endemic training points and 8% endemic training points. Sixty-eight percent of the endemic training points were classified to the high and very high susceptibility area. The moderate susceptibility class contained 24% endemic training points and 4.2% non-endemic training points.

Fig. 4.20 shows the very low and low susceptibility area covering 61.01 % of the study area and containing 100% of the non-endemic validation points. There were 20% of the endemic validation points located in the low susceptibility area. Also, 40% of the endemic validation points were categorized into high and very high susceptibility area and 40% of them were in the moderate susceptibility area. The same information is also shown in Fig 4.22. The ROC curve of this model, which produced the AUC of 0.88, is shown in Fig. 4.21.

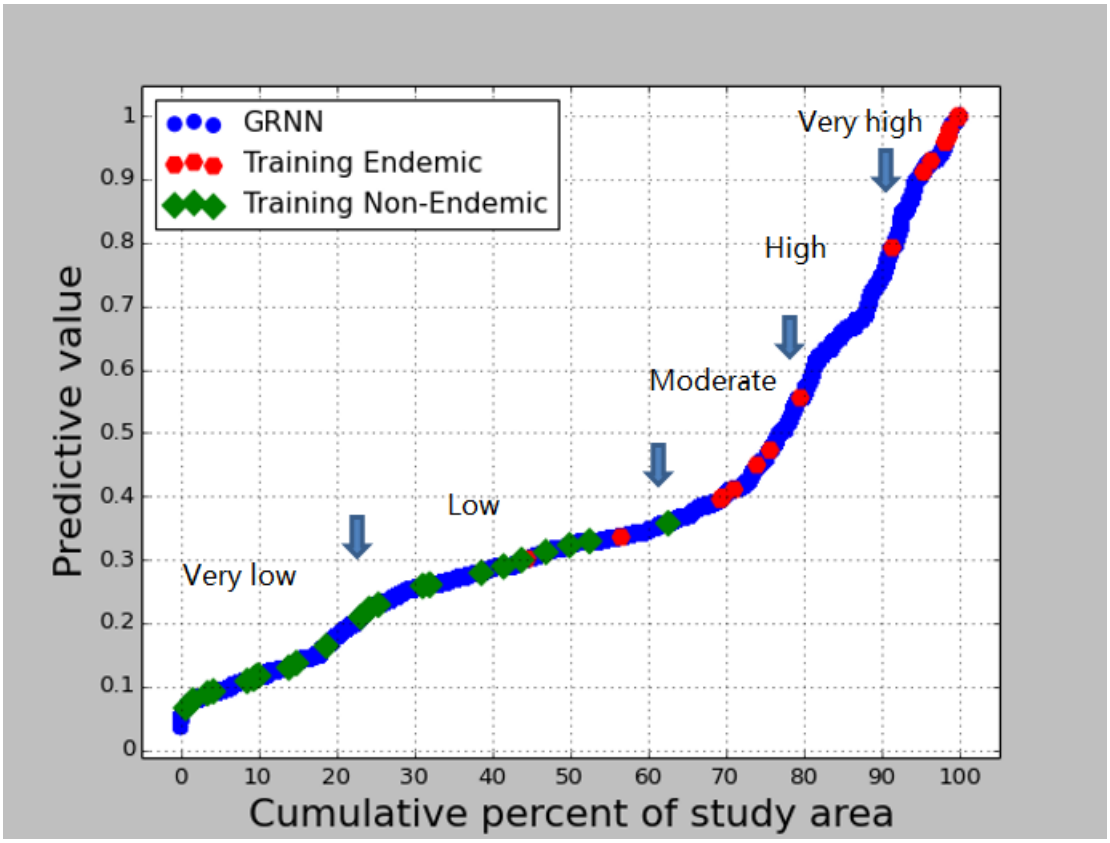


Figure 4-19.Predictive susceptibility of training points versus cumulative percent of the study area for GRNN model.

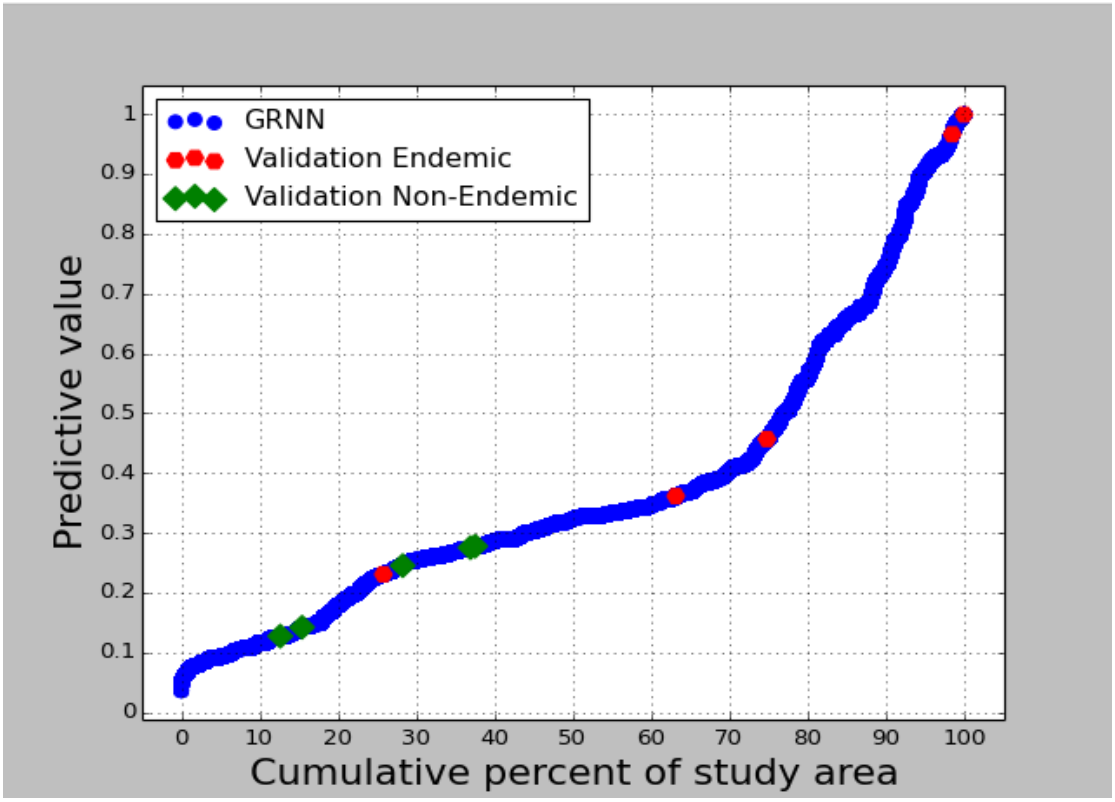


Figure 4-20.Predictive susceptibility of validation points versus cumulative percent of the study area for GRNN model.

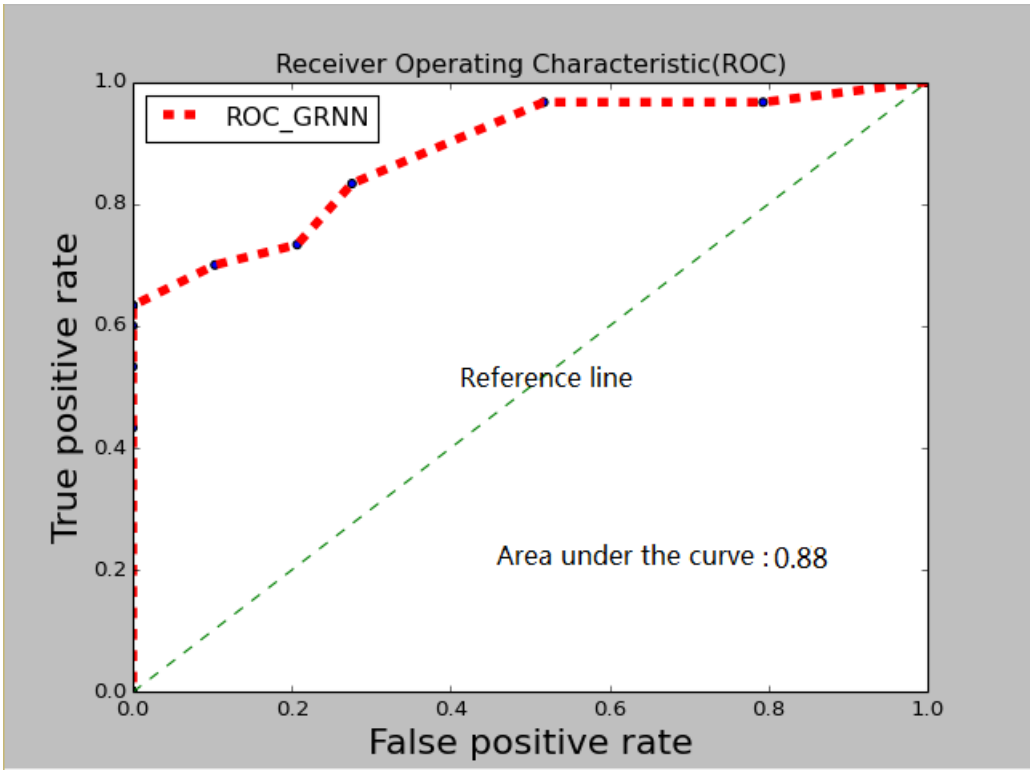


Figure 4-21.ROC for validation of GRNN model.

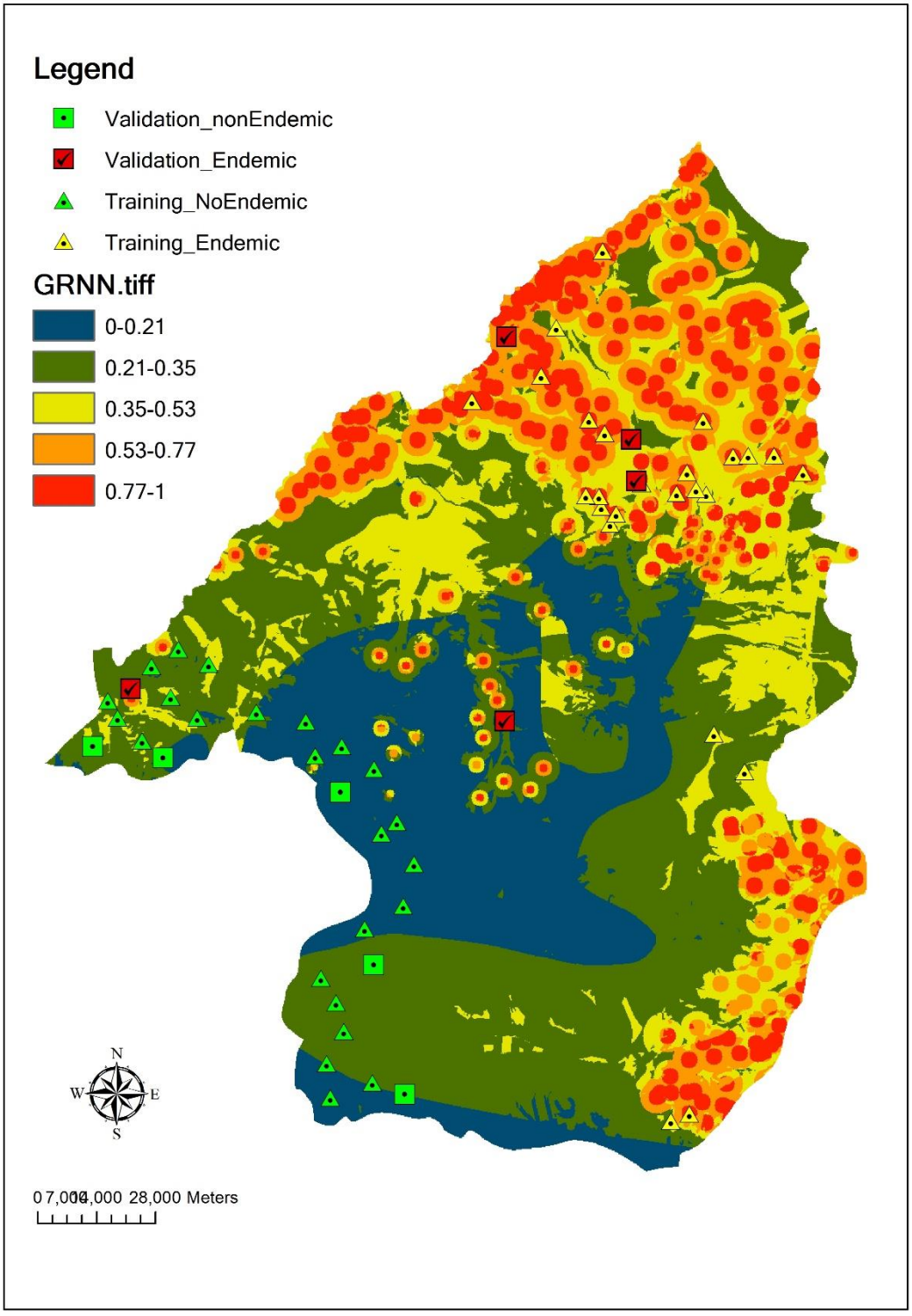


Figure 4-22. VL susceptibility map produce by GRNN.

The table 4.3 shows the AUC and accuracy generated by the four models. The BPNN generate the bset score of AUC which is 0.942 while, the GRNN is the lowest in terms of AUC which is 0.88. The AUC for LR and RBFLN are 0.899 and 0.938 respectively. Moreover, the LR, BPNN and RBFLN have the same classification accuracy of 80%. The classification accuracy of GRNN is 70%.

Table 4.3. AUC and classification accuracy generated by the four models.

Model	LR	BPNN	RRBFLN	GRNN
AUC	0.899	0.942	0.938	0.88
Accuracy	80%	80%	80%	70%

5. Discussion

5.1 Discussion of the susceptibility maps of VL

From the produced susceptibility maps of the four models, they predict that the northern and south-eastern part of the study area have high susceptibility in VL incidence. By analyzing the relationships among the evidence maps and susceptibility maps produced, the altitude and the proximity to health-centers are the two strongest factors that affect the spatial distribution of the VL. Temperature, and proximity to nomad villages also are vital in determining the occurrence of VL.

The high VL susceptibility area appears more commonly at low altitude zone and disappears in areas located of high altitude. The nomadic lifestyle associated with lower socioeconomic conditions also affects the distribution of high-risk area. The areas having lower proximity to the nomadic villages have the higher risk of incidence of VL. In addition, warmer temperature was also found to be more prone to VL. The low altitude regions, concentrated nomadic villages with warm temperature conditions, is suitable for sand fly vector hence, have the high possibility of the outbreak of VL.

The health-centers, which are helpful to disease prevention, have the effect of decreasing the risk of VL. But, it was not shown on the susceptibility maps. The proximity to rivers also influence the VL incidence. Because, the population should be concentrated in the areas closed to river and the high population density also is a factor that effects the VL incidence. While, the susceptibility maps did not show notable pattern effected by the proximity of river.

5.2 The overfitting of BPNN

As shown in 4.2, the susceptibility map generated from an overfitting BPNN model was greatly influenced by altitude and proximity to health-centers. This generated results with decreased discrimination and calibration performance presented in the

last chapter. It also influenced how the map is interpreted, since the weight of other factors are much smaller compared to the main factors.

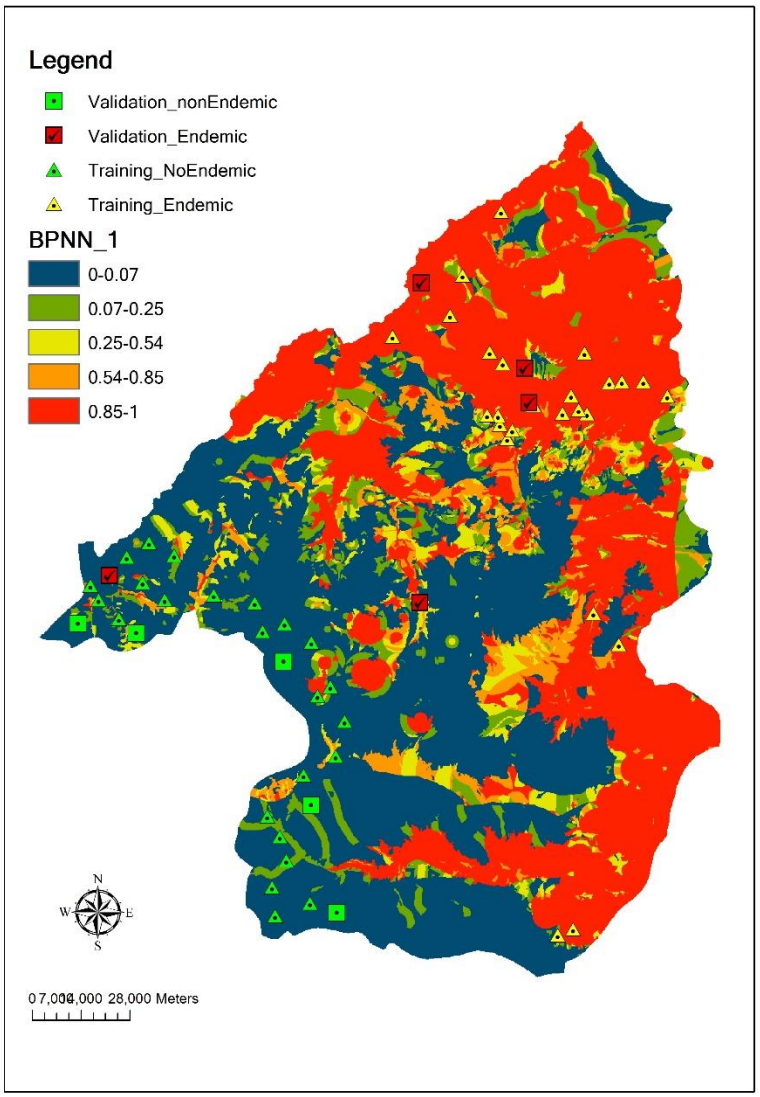


Figure 5-1.Susceptibility map produce by an overfitting BPNN model.

5.3 The setting of learning rate

As the weights of LR, BPNN and RBFLN were optimized by the algorithm of gradient decent methods. Consequently, the learning rate affects the quality of the optimal weights found. Very small learning rates result in a slow decent, leading to a longer training time than a proper learning rate. But a large learning rate may cause

an oscillation of the training SSE. It affects the quality of the optimal weights, thereby influencing the performance of the model in disease susceptibility mapping. For example, Fig. 4.3 shows the training SSE of the selected BPNN versus number of iterations under a proper learning rate. Fig. 4.4 shows training SSE of the same BPNN versus number of iterations but under a large learning rate.

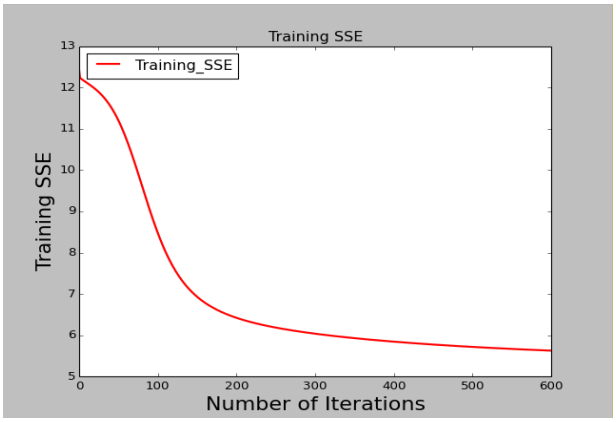


Figure 5-2. Training SSE of a proper learning rate.

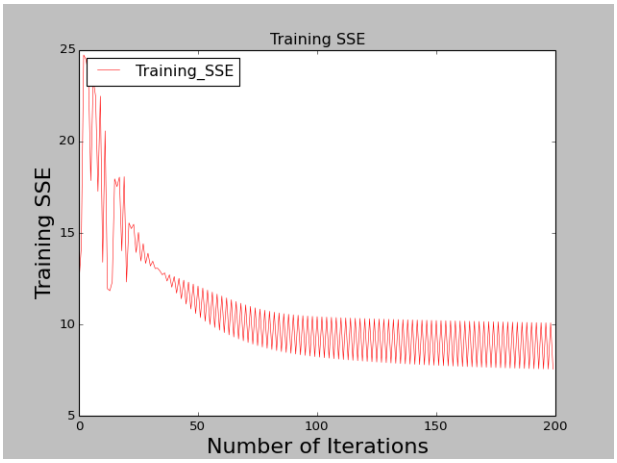


Figure 5-3. Training SSE of a large learning rate.

5.4 The performance of the four models

The LR model generated the AUC of 0.899 and classified 80% of the validation points into their proper classes. The advantages of LR are that: the structure is simple to implement but also can get a good result; and, the optimal weights obtained are interpretable, which is helpful in eliminating unrelated inputs factors and finding out the significance of the related significant ones (Yesilnacar and Topal 2005).

The BPNN generated the highest AUC of 0.942 and classified 80% of the validation points into the right class. But, BPNN has several drawbacks. Firstly, it has multiple minima, which means that the final optimal weights obtained may result from a local minimum instead of the global one. In the case study, the BPNN needed to be run several times to get the optimal weights. Secondly, the training iteration is much longer than RBFLN and GRNN. Thirdly, the selection of learning rate is based on experience, and there is no guideline for choosing the learning rate. Fourthly, the BPNN needs enough training data to estimate these parameters. But for disease susceptibility mapping, insufficiency of data commonly persists. Moreover, the optimal weights obtained cannot be interpreted (Looney 2002). Lastly, the BPNN may cram too much detail from the training data that may cause overfitting problems. Additional knowledge is needed to tackle the problem of overfitting.

The RBFLN generated the AUC of 0.938 and classified 80% of the validation points into the right class. According to (Looney 2002), the RBFLN has a unique minimum with quick training, which facilitate the process of finding the optimal weights. The only parameter to be changed is the number of nodes in the hidden layer. Moreover, the algorithm can generate reliable from small data set. But the results generated are not bounded by the minimum and maximum of the observations, which means that the final predictive score need to be normalized.

The GRNN generated the AUC of 0.88 and classified 70% of the validation points into the right class. Both of the AUC and the accuracy of GRNN is the lowest. It is because that the network is 'learn' in one pass through the data instead of the training process used in the other three methods which adjust the weights according to the training error each iteration. But, this characteristic make it generalizes from training data as soon as they are stored. It also has serval other advantages: (1) it can generate a reasonable regression surfaces based on few training data; (2) the prediction value is between the maximum and the minimum of the observations; (3) the prediction cannot converge to local minima (shown in Fig.3.20); (4) clustering methods can be used to decrease the number of nodes which can accelerate the training processes well as calculating the predictive value for each cell in the study area (Specht 1991).

6. Conclusions

Four data-driven models-LR, BPNN, RBFLN and GRNN for disease susceptibility mapping were implemented in Python. The performances of those four models were tested by a case study of VL.

In the results, the BPNN generated the best AUC which is 0.942. Meanwhile, BPNN together with RBFLN and LR correctly classified 80% of the validation data. The GRNN classified 70% of the validation data into the right class. The AUC generated by RBFLN, LR and GRNN are 0.886, 0.899 and 0.88 respectively. In the case study, all of them had the ability to highlight the main spatial distribution of the risk of VL.

Some recommendations can be made using these methods for disease susceptibility mapping. The optimal weights generated by LR are interpretable; hence, it can be used to eliminate unrelated factors and to evaluate the significance of related ones. BPNN can generate good results. But considering the long training time, the unstable results and the overfitting problems, it is recommended to use BPNN as the benchmark with which to evaluate the other results. The RBFLN can generate good result, and the training is also efficient since there is a unique minimum. Additionally, it can work well with small data sets. The RBFLN is a good artificial neural network for disease susceptibility mapping. The GRNN has a one-pass structure, which provides quick training. The clustering techniques of GRNN which decrease the number of functions in the hidden layer also make it faster to make disease maps. It is therefore suitable for quickly producing the general distribution of diseases.

Reference

Ayalew, L., and H. Yamagishi. 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology,* 65**:** 15-31. DOI: 10.1016/j.geomorph.2004.06.010

Beucher, A., P. Osterholm, A. Martinkauppi, P. Eden, and S. Frojdo. 2013. Artificial neural network for acid sulfate soil mapping: Application to the Sirppujoki River catchment area, south-western Finland. *Journal of Geochemical Exploration,* 125**:** 46-55. DOI: 10.1016/j.gexplo.2012.11.002

Brown, W. M., T. D. Gedeon, D. I. Groves, and R. G. Barnes. 2000. Artificial neural networks: a new method for mineral prospectivity mapping. *Australian Journal of Earth Sciences,* 47**:** 757-770. DOI: 10.1046/j.1440-0952.2000.00807.x

Castillo-Riquelme, M., Z. Chalabi, J. Lord, F. Guhl, D. Campbell-Lendrum, C. Davies, and J. Fox-Rushby. 2008. Modelling geographic variation in the cost-effectiveness of control policies for infectious vector diseases: The example of Chagas disease. *Journal of Health Economics,* 27**:** 405-426. DOI: 10.1016/j.jhealeco.2007.04.005

Dai, F., and C. Lee. 2002. Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. *Geomorphology,* 42**:** 213-228.

Dreiseitl, S., and L. Ohno-Machado. 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics,* 35**:** 352-359. DOI: 10.1016/s1532-0464(03)00034-0

Fung, C. C., V. Iyer, W. Brown, K. W. Wong, and Ieee. 2005. *Comparing the performance of different neural networks architectures for the prediction of mineral prospectivity*. New York: Ieee.

Hastie, T., R. Tibshirani, J. Friedman, and J. Franklin. 2005. The elements of statistical learning: data mining, inference and prediction. In *Neural networks*, 392-401.

Kumar, R., and R. Anbalagan. 2015. Remote sensing and GIS based artificial neural network system for landslide suceptibility mapping. In *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*: IEEE, 4696-4699.

Lee, S., and T. Sambath. 2006. Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. *Environmental Geology,* 50**:** 847-855. DOI: 10.1007/s00254-006-0256-7

Leonardo, L. R., B. A. Crisostomo, J. A. A. Solon, P. T. Rivera, A. B. Marcelo, and J. M. Villasper. 2007. Geographical information systems in health research and services delivery in the Philippines. *Geospatial Health,* 1**:** 147-155.

Looney, C. G. 2002. Radial basis functional link nets and fuzzy reasoning. *Neurocomputing,* 48**:** 489-

509. DOI: 10.1016/s0925-2312(01)00613-0

Peterson, A. T., and J. Shaw. 2003. Lutzomyia vectors for cutaneous leishmaniasis in Southern Brazil: ecological niche models, predicted geographic distributions, and climate change effects. *International Journal for Parasitology,* 33**:** 919-931. DOI: 10.1016/s0020-7519(03)00094-8

Porwal, A., E. Carranza, and M. Hale. 2003. Artificial neural networks for mineral-potential mapping: a case study from Aravalli Province, Western India. *Natural resources research,* 12**:** 155-171.

Rajabi, M., A. Mansourian, P. Pilesjö, and A. Bazmani. 2014. Environmental modelling of visceral leishmaniasis by susceptibility-mapping using neural networks: a case study in north-western Iran. *2014,* 9**:** 13. DOI: 10.4081/gh.2014.15

Salahi-Moghaddam, A., M. Mohebali, A. Moshfae, M. Habibi, and Z. Zarei. 2010. Ecological study and risk mapping of visceral leishmaniasis in an endemic area of Iran based on a geographical information systems approach. *Geospatial Health,* 5**:** 71-77.

Seid, A., E. Gadisa, T. Tsegaw, A. Abera, A. Teshome, A. Mulugeta, M. Herrero, D. Argaw, et al. 2014. Risk map for cutaneous leishmaniasis in Ethiopia based on environmental factors as revealed by geographical information systems and statistics. *Geospatial Health,* 8**:** 377-387.

Singer, D. A., and R. Kouda. 1996. Application of a feedforward neural network in the search for kuroho deposits in the Hokuroku District, Japan. *Mathematical Geology,* 28**:** 1017-1023. DOI: 10.1007/bf02068587

Specht, D. F. 1991. A general regression neural network. *Neural Networks, IEEE Transactions on,* 2**:** 568-576.

Tsegaw, T., E. Gadisa, A. Seid, A. Abera, A. Teshome, A. Mulugeta, M. Herrero, D. Argaw, et al. 2013. Identification of environmental parameters and risk mapping of visceral leishmaniasis in Ethiopia by using geographical information systems and a statistical approach. *Geospatial Health,* 7**:** 299-308.

WHO. 2016a. Vector-borne diseases. Retrieved 3rd, May 2016, from http://www.who.int/mediacentre/factsheets/fs387/en/

WHO. 2016b. WHO fact sheets, Leishmaniasis. Retrieved 25 April 2016, from http://www.who.int/mediacentre/factsheets/fs375/en/.

Yesilnacar, E., and T. Topal. 2005. Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). *Engineering Geology,* 79**:** 251-266. DOI: 10.1016/j.enggeo.2005.02.002

Yilmaz, I. 2009. Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: A case study from Kat landslides (Tokat-Turkey). *Computers & Geosciences,* 35**:** 1125-1138. DOI: 10.1016/j.cageo.2008.08.007

Department of Physical Geography and Ecosystem Science, Lund University

Lund University GEM thesis series are master theses written by students of the international master program on Geo-information Science and Earth Observation for Environmental Modelling and Management (GEM). The program is a cooperation of EU universities in Iceland, the Netherlands, Poland, Sweden and UK, as well a partner university in Australia. In this series only master thesis are included of students that performed their project at Lund University. Other theses of this program are available from the ITC, the Netherlands (www.gem-msc.org or www.itc.nl).

The student thesis reports are available at the Geo-Library, Department of Physical Geography and Ecosystem Science, University of Lund, Sölvegatan 12, S-223 62 Lund, Sweden. Report series started 2013. The complete list and electronic versions are also electronic available at the LUP student papers (https://lup.lub.lu.se/student-papers/search/) and through the Geo-library (www.geobib.lu.se).

1 Soheila Youneszadeh Jalili (2013) The effect of land use on land surface temperature in the Netherlands
2 Oskar Löfgren (2013) Using Worldview-2 satellite imagery to detect indicators of high species diversity in grasslands
3 Yang Zhou (2013) Inter-annual memory effects between Soil Moisture and NDVI in the Sahel
4 Efren Lopez Blanco (2014) Assessing the potential of embedding vegetation dynamics into a fire behaviour model: LPJ-GUESS-FARSITE
5 Anna Movsisyan (2014) Climate change impact on water and temperature conditions of forest soils: A case study related to the Swedish forestry sector
6 Liliana Carolina Castillo Villamor (2015) Technical assessment of GeoSUR and comparison with INSPIRE experience in the context of an environmental vulnerability analysis using GeoSUR data
7 Hossein Maazallahi (2015) Switching to the "Golden Age of Natural Gas" with a Focus on Shale Gas Exploitation: A Possible Bridge to Mitigate Climate Change?
8 Mohan Dev Joshi (2015) Impacts of Climate Change on Abies spectabilis: An approach integrating Maxent Model (MAXent) and Dynamic Vegetation Model (LPJ-GUESS)
9 Altaaf Mechiche-Alami (2015) Modelling future wheat yields in Spain with LPJ-GUESS and assessing the impacts of earlier planting dates
10 Koffi Unwana Saturday (2015) Petroleum activities, wetland utilization and livelihood changes in Southern Akwa Ibom State, Nigeria: 2003-2015
11 José Ignacio Díaz González (2016) Multi-objective optimisation algorithms for GIS-based multi-criteria decision analysis: an application for evacuation planning
12 Gunjan Sharma (2016) Land surface phenology as an indicator of performance of conservation policies like Natura2000
13 Chao Yang (2016) A Comparison of Four Methods of Diseases Mapping