Predictive risk mapping of human leptospirosis using support vector machine classification and multilayer perceptron neural network

Mehrdad Ahangarcani,[1] Mahdi Farnaghi,[1,2] Mohammad Reza Shirzadi,[3] Petter Pilesjö,[2,4] Ali Mansourian[2,4]

[1]Faculty of Geodesy and Geomatics Engineering, K. N. Toosi University of Technology, Tehran, Iran;
[2]GIS Center, Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden;
[3]Center for Disease Control (CDC), Ministry of Health and Medical Education, Tehran, Iran;
[4]Center for Middle-Eastern Studies, Lund University, Lund, Sweden

Abstract

Leptospirosis is a zoonotic disease found wherever human is in direct or indirect contact with contaminated water and environment. Considering the increasing number of cases of this disease in the northern part of Iran, identifying areas characterized by high disease incidence risk can help policy-makers develop strategies to prevent its further spread. This study presents an approach for generating predictive risk maps of leptospirosis using spatial statistics, environmental variables and machine learning. Moran's *I* demonstrated that the distribution of leptospirosis cases in the study area in Iran was highly clustered. Pearson's correlation analysis was conducted to examine the type and strength of relationships between climate and topographical factors and incidence of the disease. To handle the complex and nonlinear problems involved, machine learning based on the support vector machine classification algorithm and multilayer perceptron neural network was exploited to generate annual and monthly predictive risk maps of leptospirosis distribution. Performance of both models was evaluated using receiver operating characteristic curve and Kappa coefficient. The output results demonstrated that both models are adequate for the prediction of the probability of leptospirosis incidence.

Correspondence: Mahdi Farnaghi, GIS Centre, Department of Physical Geography and Ecosystem Science, Lund University, Sölvegatan 12, SE-223 62 Lund, Sweden.
E-mail: mahdi.farnaghi@nateko.lu.se

Key words: Leptospirosis; Geographical information systems; Support vector machine; Multilayer perceptron neural network; Geostatistics; Iran.

Acknowledgments: the authors would like to express their sincere thanks to the personnel of the Center for Disease Control and Prevention in Ministry of Health and Medical Education of Iran who provided them with the leptospirosis data. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

See online Appendix for additional Tables and Figures.

Contributions: MA and MF conceived and designed the study and related experiments. MRS provided the data. MA implemented the experiments and statistical analysis. MA and MF prepared the original draft. MF was the supervisor and edited the manuscript. AM and PP reviewed the manuscript and improved the content.

Conflict of interest: the authors declare no potential conflict of interest.

Funding: none.

Received for publication: 28 May 2018.
Revision received: 29 November 2018.
Accepted for publication: 4 December 2018.

Geospatial Health 2019; 14:711
doi:10.4081/gh.2019.711

Introduction

Human leptospirosis is one of the most widespread disease among humans and animals (Vega-Corredor and Opadeyi, 2014). Caused by the bacterium *Leptospira*, this zoonotic disease has become an increasing public health problem worldwide (Adler, 2015). It is found wherever humans are in direct or indirect contact with water, damp soil and similar environments contaminated by infected blood, urine or tissue of carrier animals and is transmitted to humans through mucous membranes or cuts and abrasions on the skin (Vega-Corredor and Opadeyi, 2014). The infectious source consists mostly of rats, but also other wild animals as well as domestic ones can be carriers (Lau *et al.*, 2010). The disease is chiefly found in tropical, subtropical, hot and humid areas with high rainfall (Terpstra, 2003; Honarmand *et al.*, 2007; Rafiei *et al.*, 2012; Vega-Corredor and Opadeyi, 2014). Nowadays, because of people commonly commuting between rural and urban areas, keeping pet animals at home and particularly with respect to job-related activities, such as butchering and working in slaughterhouses, leptospirosis has changed from a villagers' traditional disease, mostly seen among farmers and fishermen, to an epidemic disease also in urban communities with inappropriate health conditions (Rafiei *et al.*, 2012). Although the number of human cases is not known precisely, available reports show that the incidence rate range from 0.1-1 per 100,000 persons per year in temperate areas to 10-100 per 100,000 persons per year in the humid tropics

(Terpstra, 2003; Vega-Corredor and Opadeyi, 2014). In recent years, the disease has been recorded in Iran, especially in the northern provinces with mild and wet weather conditions during the warm season (Rafiei *et al.*, 2012). In these provinces, rice cultivation is the dominant activity. Being exposed to contaminated water in paddy fields which are mostly irrigated by surface and stagnant waters can be a contributing factor to the increased rate of leptospirosis cases seen in the northern provinces (Honarmand *et al.*, 2007; Rafiei *et al.*, 2012).

Considering the high incidence of leptospirosis in areas with hot and humid weather conditions, predictive risk maps based on these environmental variables could help public health managers to develop controlling strategies and prevent further spread of the disease. Although there have been a plethora of studies about risk mapping and identifying critical areas of infection diseases (Ali *et al.*, 2002; Rajabi *et al.*, 2012; Rajabi *et al.*, 2014; Baggenstos *et al.*, 2016; Rajabi *et al.*, 2016; Ramezankhani *et al.*, 2017a; Ramezankhani *et al.*, 2017b; Raei *et al.*, 2018), so far, to the best of our knowledge, researchers have solely investigated the relationship among leptospirosis incidence and the environmental factors (Barcellos and Sabroza, 2001; Lau *et al.*, 2012; Gracie *et al.*, 2014; Vega-Corredor and Opadeyi, 2014; Mohammadinia *et al.*, 2015) without dealing with spatio-temporal risk predictions. Annual predictive risk maps of leptospirosis distribution, generated by a spatial-temporal method, can identify high-risk areas of the disease. Such map could be an essential tool for public health policy-makers to take pre-emptive action in vulnerable areas. Also, given the fact that high incidence of leptospirosis occurs in certain months of the year, monthly predictive risk maps of its distribution could be useful for the detection of critical places and potential disease progress in the coming months. Having this information, health care resources and facilities can be allocated more precisely. The main objectives of this study were to analyze the spatial distribution as well as possible clusters of leptospirosis in the northern provinces of Iran, explore the association between effective factors and disease incidence, and present a solution to generate the annual and monthly predictive risk maps of leptospirosis distribution. In this regard, geostatistical methods were used to detect the distribution pattern of leptospirosis in the study area. Pearson's correlation coefficient was exploited to investigate the relationships between effective factors and incidence. In order to generate annual and monthly predictive risk maps of the distribution of leptospirosis, a solution based on support vector machine (SVM) learning and multilayer perceptron (MLP) machine learning algorithms, which are well known for their ability to handle complex and nonlinear problems (Ahmad *et al.*, 2014; Pezeshki *et al.*, 2016), were developed. As data-driven algorithms, SVM and MLP can infer the intrinsic relationships among input and output parameters without the requirement for explicit definition of the impacting processes (Hippert *et al.*, 2001). Due to the complexity, nonlinearity and diversity of the affecting factors, SVM and MLP were exploited to generate the predictive risk maps of leptospirosis in this study. The performance of both algorithms was evaluated using receiver operating characteristic (ROC) curve and Kappa coefficient measures.

Materials and Methods

Study area and data

The study area includes Gilan, Mazandaran and Golestan provinces of Iran with a joint population of 7,331,831 according to 2011 census data. The area consists of 119 districts that cover an area of 58,250 km^2 and is located in the North of the country (Figure 1). These areas have an average annual rainfall of 948 mm. Almost half of the areas covered are dedicated to farming. Located along the Caspian coast and Alborz Mountains, these provinces

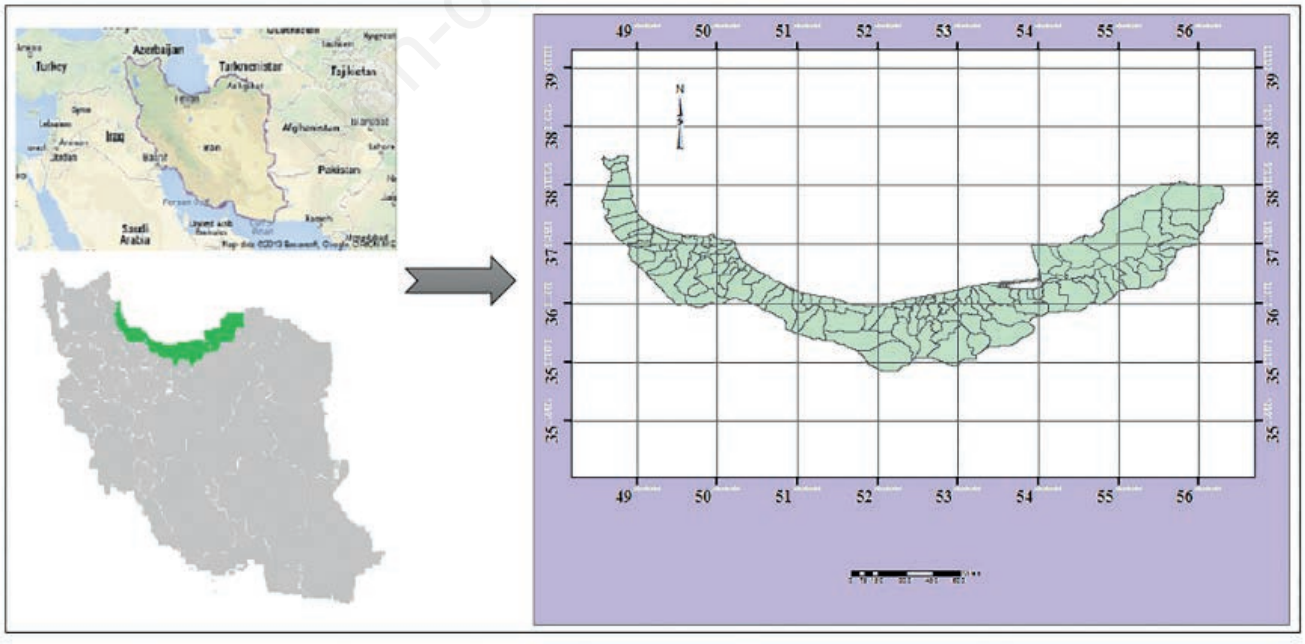

Figure 1. Northern provinces of Iran and their districts.

have a temperate and humid climate.

A longitudinal study of human leptospirosis was performed by the Center for Disease Control and Prevention of the Ministry of Health during the period of January 2009 to December 2014 at the district level throughout the study areas in which a dataset of 1,863 cases was collected. The data include monthly reported cases and their occurrence location at the district level. To ensure data reliability, detection of antibody to *Leptospira* was carried out using the enzyme-linked immunosorbent assay test, microscopic agglutination test and indirect fluorescent antibody test were applied on all the recorded cases, from which valid cases were selected.

To specify the effective factors needed to generate the predictive risk map of leptospirosis, the association between incidence and various factors mentioned in the literature were explored. In this regard, climate and topographical factors were deduced to have a profound impact on the occurrence of leptospirosis cases (Barcellos and Sabroza, 2001; Lau *et al.*, 2012; Vega-Corredor and Opadeyi, 2014). Also, knowing about incidence rate of this infection in the previous months and years can help to better predict the incidence rate in future. To reach our objectives monthly and annual data would be required. To that end, data were collected for the period January 2009 to December 2014 from various sources (Table 1). Annual and monthly evidence maps corresponding to each factor were produced using geographical information systems software. For example, Figure A1 in the Appendix shows the annual and monthly standardized incidence rate of leptospirosis at the district level in the study area where the areas vulnerable to leptospirosis were separated into 5 distinct zones, namely: very low, low, medium, high and very high.

The support vector machine

Developed by Vapnik (1998), SVM is a classification algorithm that has its root in statistical learning theory. It is a supervised binary classification method (Burges, 1998). However, one-against-all (OAA) and one-against-one (OAO) techniques can be exploited to use SVM as a multi-class classifier (Duan and Keerthi, 2005). Assuming a linear separability of classes, SVM classifier tries to find the optimal separating hyperplane that maximizes the distance (margin) between the closest samples of classes. Dealing with nonlinear problems, the SVM classifier maps the input data into a space with higher dimensions using kernel functions, so that the data can be linearly separated in this new space (Ben-Hur and Weston, 2010).

Having a dataset, $\{x_i, y_i \mid i = 1,2,3 \ldots, l\}$ where $x_i \in R^d$ is the input vector and $y_i \in \{-1, +1\}$ the corresponding label, a SVM classification function can be formulated by Equation 1 (Burges, 1998).

$$f(x) = sign\left(\sum_{i=1}^{l} \alpha_i y_i \, k(x_i, x) + b\right) \quad \text{Eq. 1}$$

where α_i is a Lagrange's multiplier, $k(x_i, x)$ the kernel function, b the bias term and C the regularization constant that controls the trade-offs between maximizing the margin and decreasing the errors. Having a training dataset, the Lagrange multipliers are obtained by solving Equation 2.

$$Maximize \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \, k(x_i, x_j)$$
$$subject\ to \sum_{i=1}^{l} \alpha_i y_i = 0 \ and \ 0 \le \alpha_i \le C \ for \ all \ i \quad \text{Eq. 2}$$

The closest data points to the hyperplane with $0 < \alpha_i \le C$ are considered the support vectors. Having the Lagrange multipliers and the support vectors, each new data sample can be classified using Equation 1.

The multilayer perceptron neural network

Artificial neural networks (ANNs) mimic the function of the human brain and can be defined as simplified mathematical models which can be trained to learn (Anderson, 1995). The fundamental elements of ANN are neurons that are organized in layers and receive multiple signals, combine and modify them to transmit the result to other neurons (Hagan *et al.*, 1996). ANN is able to solve linear and nonlinear problems using linear and nonlinear activation functions (Anderson, 1995).

In this study, we used a particular type of ANN, called MLP

Table 1. Data used in leptospirosis modeling.

Data	Source	Evidence map
Climate	Meteorological organization of Iran	Sum precipitation Mean temperature Mean humidity Number of days below 0ºC
Topographical	Digital Elevation Models derived from Shuttle Radar Topography Mission satellite images with a spatial resolution of 30 m	Mean altitude Slope
	Moderate Resolution Imaging Spectroradiometer satellite images with a spatial resolution of 250 m	Land cover
Leptospirosis incidence	Center for Disease Control and Prevention of Ministry of Health of Iran of Ministry of Health of Iran	Standardized Incidence Rate $= \frac{\text{Total cases of leptospirosis}}{\text{Population of the region}} * 100000$
Census	Statistical Center of Iran	

(Hornik *et al.*, 1989). An MLP model consists of three layer types, including one input layer, one (or more) hidden layers and one output layer. The neurons in each layer are fully connected to the neurons in the next layer, but not connected to the neurons in their own layer. An MLP is a classifier that receives the data by the input layer and transforms it into a new space where it becomes linearly separable using non-linear transformation. The hidden layer performs this transformation, while the output layer classifies the data in the new space.

Equation 3represents a formal one-hidden-layer MLP.

$$f(x) = G\left(b^{(2)} + W^{(2)}s\left(b^{(1)} + W^{(1)}x\right)\right) \quad \text{Eq. 3}$$

where $b^{(1)}$ and $b^{(2)}$ are the bias vectors, $W^{(1)}$ and $W^{(2)}$ the weight matrices with G and s the activation functions. Having a train dataset, an MLP model is trained by a backpropagation learning process (Williams and Hinton, 1986), through which the model parameters ($b^{(1)}$, $b^{(2)}$, $W^{(1)}$ and$W^{(2)}$) are determined. Having the model parameters, each new data sample can be classified using Equation 3.

Research methodology

The goals of this study were to analyze the spatial distribution of leptospirosis in the study area, explore the associations among effective factors and diseases incidences proposing a solution to generate the annual and monthly predictive risk map of leptospirosis distribution. In order to achieve these goals, the following four significant steps were conducted: i) Moran's *I* (Moran, 1948) was used to detect the annual and monthly distribution pattern of leptospirosis cases. Additionally, Local Moran's *I* (Anselin, 1995; Caldas de Castro and Singer, 2006) was used to detect spatial leptospirosis clusters (hotspots) in annual and monthly datasets; ii) Pearson's correlation coefficient (Benesty *et al.*, 2009) was exploited to examine the type and strength of the associations among climate and topographical factors with annual and monthly incidence of leptospirosis; iii) the SVM classifier and MLP neural network were used to map leptospirosis and affecting factors and generate the annual and monthly predictive risk map. For this purpose, the SVM and MLP models were built so that they would be able to predict the incidence by receiving climate and topographical factors. The two models were implemented by Java using the jKernelMachines (https://github.com/davidpicard/jkernelmachines) library and Neuroph (http://neuroph.sourceforge.net/) library; and iv) the performance of the SVM and MLP models were compared using ROC curve (Bradley, 1997) and Kappa coefficient (Carletta, 1996). The ROC curve is a plot of sensitivity (true positive rate) on the y-axis and specificity (false positive rate) on the x-axis. For a particular model, the area under the curve (AUC) can be used as a measure of classification performance (Brown *et al.*, 2003; Stevens *et al.*, 2013) so that an AUC close to 0.5 indicates that the result of the model is random, while an AUC close to 1 shows that the model is able to acceptably separate classes (Brown *et al.*, 2003). Additionally, Kappa coefficient measures the similarity of observed and predicted values from the spatial distribution point of view so that a Kappa value of 1 represents a perfect agreement between the observed and the predicted maps, while a Kappa value of 0 represents none (Sousa *et al.*, 2002).

Particularly in the third step, the SVM and MLP models were built, calibrated, trained and tested with the aim of both models to estimate the function of Equation 4.

$$y = f(\vec{X}) \quad \text{Eq. 4}$$

where y denotes class label of leptospirosis incidence rate at time t; and $(\vec{X})$ represents a vector of effective factors including altitude, slope, land cover, average temperature, average humidity, rainfall and the number of days below 0ºC at t and incidence rate of leptospirosis at t-1. In the annual mode, t refers to the current year and t-1 indicates the previous year. In the same way, in the monthly mode, t refers to the current month and t-1 the previous month. These models are used for both annual and monthly predictions.

To calibrate the SVM model and acquire the accurate results, it is necessary to determine the kernel function and its parameters including the *C* and γ parameters. Grid-search and cross-validation methods (Hsu *et al.*, 2003) have been used to determine the relevant parameters. Kernel functions that are most widely used in various studies are linear, Gaussian (radial basis function, RBF) and polynomial (Hsu *et al.*, 2003). In this study, the performance of linear, RBF and polynomial kernel functions were compared with each other. The RBF kernel function offered the highest precision. Best values of *C* and γ parameters were obtained 2 and 0.0023, respectively (Table A1 in the Appendix shows the performance of different combinations of kernel functions along with the C and γ parameters). Due to the low number of classes (5 classes), OAA strategy was used to be able to exploit SVM as a multiclass classifier.

In order to calibrate the MLP model, it is necessary to determine activity functions and the number of neurons in the hidden layer. Like the SVM model, grid-search and cross-validation methods were used to determine the relevant parameters of the MLP model. Various activity functions such as linear, threshold, sigmoid and hyperbolic tangent (Tanh) have been used in the literature. In this study, due to the complexity and non-linearity of the problem, different combinations of activity functions and different numbers of neurons in the hidden layer were compared with each other (Table A2 in the Appendix shows the comparison results). The comparison showed that sigmoid activity function along with a hidden layer of 8 neurons offers the highest precision and this structure was selected for the MLP model.

After calibration of SVM and MLP models, 80% of the dataset, pertaining to 2009-2013 (years 2009-2013 in annual prediction and January 2009 to December 2013 in monthly prediction), which included both class labels and factors, were used to train the models. The remaining 20% of the data, pertained to 2014 (the year 2014 in annual prediction and May 2014 to September 2014 in monthly prediction) were afterward used to generate the predictive risk map of leptospirosis distribution and test the accuracy of the trained SVM and MLP models. The trained SVM and MLP models should then be able to predict the incidence rate of leptospirosis at the district level in the future.

Results

Cluster detection

Results of applying Moran's *I* on leptospirosis incidence maps indicated that the distribution of cases in the northern provinces of Iran for both annual and monthly observations are highly clustered (Table A3 and Table A4 in the Appendix show that Moran's *I* indexes reject the null hypothesis and therefore the

distributions of the disease in the study area are clustered).

Figure 2 shows the results of Local Moran's *I*, identified annual spatial clusters of leptospirosis cases with high incidence rates in the study areas where the areas marked HH, represent hotspots.

As it can be seen in Figure 3, the most significant difference in the monthly leptospirosis clustering was seen in August and May. This may be due to climate difference as well as changes in the lifestyle of inhabitants between these two months.

Association among effective factors and leptospirosis incidence

Table 2 shows the annual relationship between climate and topographical factors and leptospirosis incidence. Based on the results of Pearson's correlation, positive correlations were observed between the number of disease cases with average humidity, average temperature, rainfall and leptospirosis incidence in the previous year. Also, negative correlations were seen between the incidence with altitude, slope, land cover and number of days

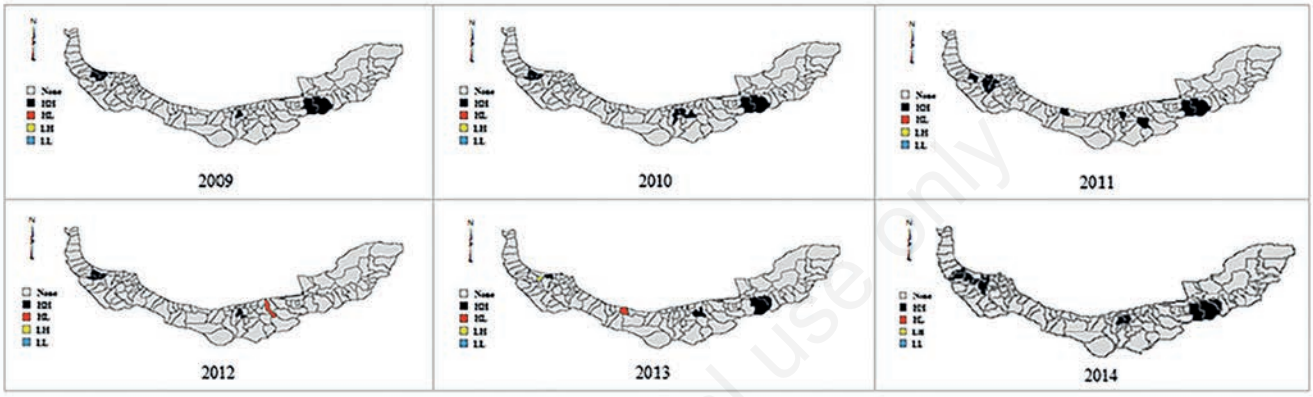

Figure 2. Position of leptospirosis clusters detected using the local Moran's *I* in the annual survey.HH, high-high spatial association; HL, high-low spatial association; LH, low-high spatial association; LL, low-low spatial association.

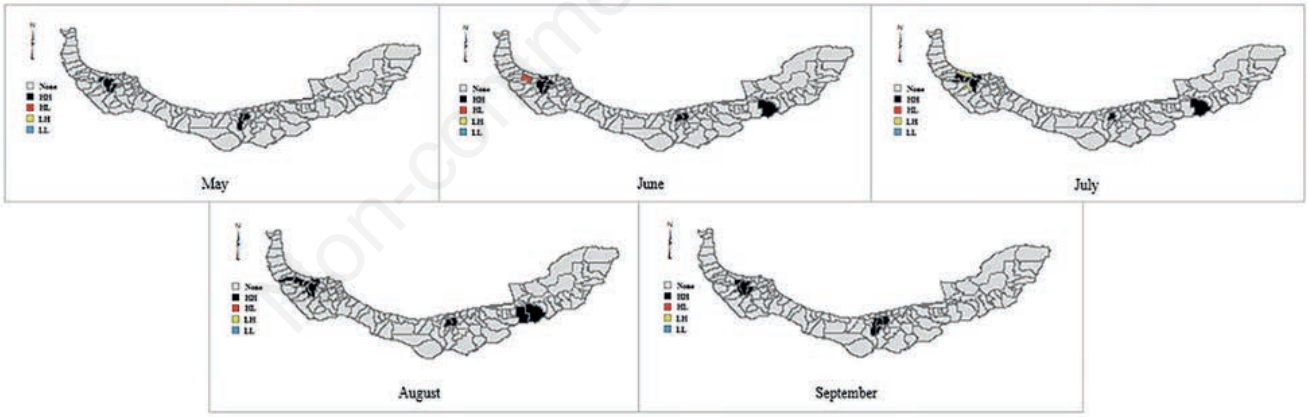

Figure 3. Position of leptospirosis clusters detected using the local Moran's *I* in the monthly survey (2014). HH, high-high spatial association; HL, high-low spatial association; LH, low-high spatial association; LL, low-low spatial association.

Table 2. Results of Pearson correlation analysis between climate and topographical factors and annual leptospirosis incidence.

Dependent variable	Independent variables							
	Altitude	Slope	Land cover	Average temp.	Average humidity	Rain-fall	Days <0°C	Incidence of leptospirosis previous year
Incidence of leptospirosis	-0.36	-0.27	-0.18	0.41	0.45	0.39	-0.11	0.06

Correlation significant at the P≤0.01 level.

below 0ºC. Among the effective factors, regardless of negative or positive signs, average humidity, average temperature, rainfall, altitude, and slope factors showed the highest correlation with the leptospirosis incidence, respectively. Whereas the incidence in the previous year, the number of days below 0ºC and the land cover showed the least such.

Table 3 shows the monthly relationship between effective factors and leptospirosis incidence in the study area. The monthly results highly match the obtained results of the annual analysis.

Predictive risk map of leptospirosis

Figures 4 and 5 show the results of using the SVM and MLP models to create annual and monthly predictive risk maps of leptospirosis, respectively.

Additionally, Tables 4 and 5 indicate the results of applying the evaluation measures including ROC curve and Kappa coefficient on the annual and monthly data.

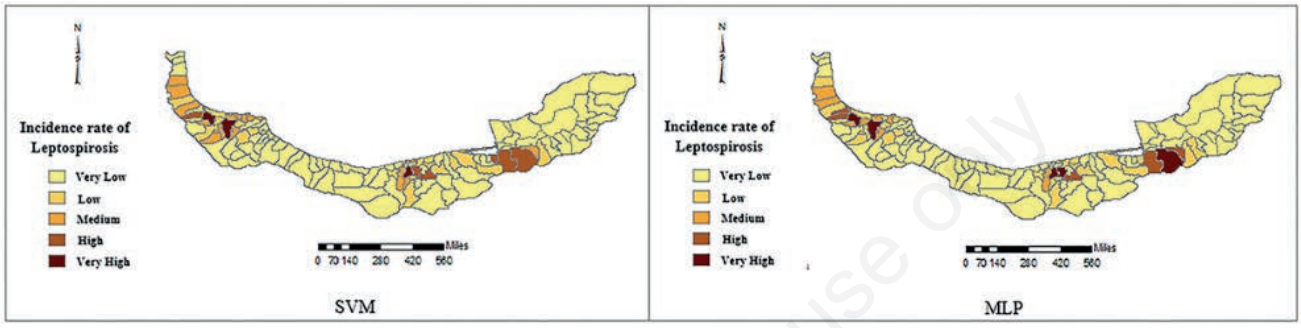

Figure 4. Annual predicted incidence rate of leptospirosis in 2014. SVM, support vector machine; MLP, multilayer perceptron.

Table 3. Results of Pearson correlation analysis between climate and topographical factors and monthly leptospirosis incidence.

Dependent variable	Independent variables							
Leptospirosis incidence	Altitude	Slope	Land cover	Average temp.	Average humidity	Rain-fall	Days <0ºC	Incidence of leptospirosis previous year
May	-0.397	-0.241	-0.228	0.524	0.586	0.493	-0.105	0.052
June	-0.428	-0.264	-0.217	0.649	0.711	0.469	-0.094	0.063
July	-0.401	-0.257	-0.119	0.703	0.755	0.441	-0.076	0.072
August	-0.397	-0.248	-0.116	0.834	0.862	0.497	-0.072	0.088
September	-0.405	-0.252	-0.225	0.543	0.561	0.514	-0.083	0.067

Correlation significant at the P≤0.01 level.

Table 4. Kappa coefficient and area under the curve (AUC) of annual prediction.

Method	Number being correct	Number being wrong	Level of being correct (%)	Kappa coefficient (%)	AUC
Support vector machine	103	16	86.55	85.13	0.8548
Multilayer perceptron	101	18	84.87	83.64	0.8336

Table 5. Kappa coefficient and area under the curve (AUC) of monthly prediction.

Month	Number being correct (SVM)	Number being correct (MLP)	Level of being correct (SVM)	Level of being correct (MLP)	Kappa coefficient (SVM)	Kappa coefficient (MLP)	AUC (SVM)	AUC (MLP)
May	105	104	88.23	87.39	88.14	86.24	0.8802	0.8652
June	105	103	88.23	86.55	87.91	86.07	0.8769	0.8618
July	103	102	86.55	85.71	85.67	83.92	0.8518	0.8363
August	105	103	88.23	86.55	87.96	86.31	0.8732	0.8637
Sept.	104	102	87.39	85.71	87.03	84.43	0.8674	0.8430

SVM, support vector machine; MLP, multilayer perceptron.

Discussion

Public health policy-makers and managers are often interested in visualizing and studying the distribution of diseases as it can help public health policy-makers and managers to determine the priority areas for the allocation of budget, personnel, and equipment to mitigate and respond to emergencies. As expressed by Moore and Carpenter (1999) and Mollalo *et al.* (2015), visualization alone cannot indicate whether an aggregation of incidence cases is due to a large number of cases or a large population in an

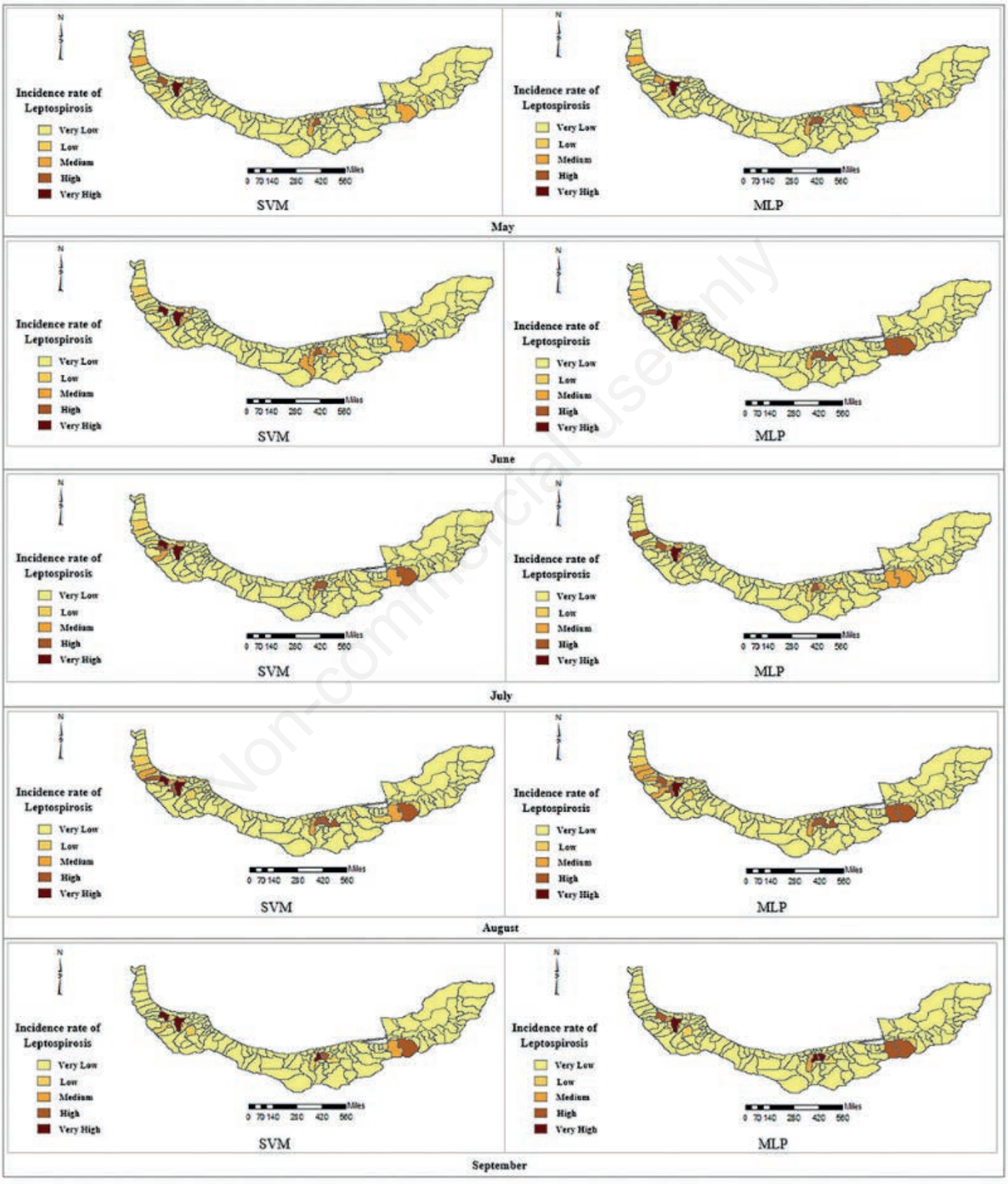

Figure 5. Monthly predicted incidence rate of leptospirosis in 2014. SVM, support vector machine; MLP, multilayer perceptron.

area. With this in mind, consideration of the population effect is inevitable. Consequently, in this study, instead of using the number of leptospirosis cases, the standardized incidence rate (according to Table 1) was used in geostatistical and statistical analyses as well as generating predictive risk maps.

Moran's *I* illustrates a clustering distribution pattern of leptospirosis in the northern part of Iran. In 2014, in particular, the values of Global Moran's *I* in May (I=0.19, Z=3.57, P<0.01), June (I=0.37, Z=6.60, P<0.01), July (I=0.15, Z=2.87, P<0.01), August (I=0.17, Z=3.19, P<0.01) and September (I=0.21, Z=4.13, P<0.01) show that the distribution patterns were not spatially random. Given the results of local Moran's *I* (Figures 2 and 3), the most likely spatial clusters were seen in the central districts of Gilan Province, the north-eastern districts of Mazandaran Province and the western districts of Golestan Province. These are also the areas where agricultural and paddy activities are concentrated and hence can be the main reason for the formation of the clusters.

The Pearson's correlation coefficient analysis, presented in Tables 2 and 3, as well as a closer look at the overlap of the locations of the leptospirosis clusters with climatic and topographical conditions, show that high-risk clusters were located in flat and low slope areas with hot and humid weather along with heavy rainfall and sparse land cover. This landscape is typical for most agricultural and paddy fields in the northern part of Iran, which often leads to the accumulation of surface and stagnant waters. In this situation, higher evaporation from surface water and intensified precipitation, the probability of flooding increases, which prevents absorption of the contagion in the soil and promoting transmission into human habitats. As the temperature rises, activities such as agriculture and farming, fishing, water entertainments, *etc*. increase, which lead to more contact of humans and animals with *e.g*. infected waters eventually resulting in a surge in leptospirosis occurrence. The outcomes of this study with respect to the climatic and topographical conditions discussed above are in line with Karande *et al.* (2003), Honarmand *et al.* (2007), Rafiei *et al.* (2012) and Vega-Corredor and Opadeyi (2014). Also, the Pearson's correlation coefficient analysis indicated that leptospirosis incidence in the previous month or year has a low impact on the current leptospirosis occurrence, meaning that transmission from human to humans is rare, which is consistent with the findings of Terpstra (2003).

The developed SVM and MLP models were able to forecast risk areas of leptospirosis with acceptable accuracy. Comparison of prediction and observation shows that predictive risk maps acceptably match the observed maps. The AUC values close to 1 for ROC curves and Kappa coefficient values close to 1 demonstrate that both models were able to predict the annual and monthly leptospirosis incidence (Tables 4 and 5). The output predictive risk maps can be used to improve the decision-making process in public health management and provide essential guidelines for policy-makers to both monitor leptospirosis distribution and predict its incidence. With a precise and location-based prediction of leptospirosis incidence, public health managers at regional and local levels can allocate budget, personnel and resources more efficiently and effectively concentrate on the areas with the highest risk. Additionally, these output maps can be used as reconnaissance guides for decision makers to effectively peruse controlling strategies such as teaching and preventive measures that can reduce the number of leptospirosis cases in high-risk areas. Also, policy-makers can run these predictive models based on different scenarios of climate and topography and develop guidelines and solutions for preventing further spread of leptospirosis and respond to disease incidence in different conditions.

Conclusions

In this study, the geostatistical analysis indicated that the distribution of leptospirosis cases in the study area was clustered. Pearson's correlation analysis showed that climate and topographical factors significantly affect the spatial distribution of this infection. The two powerful machine learning techniques MLP and SVM were shown to accurately generate annual and monthly predictive risk map of leptospirosis distribution.

As a future work, we suggest that the effect of other environmental and socio-economic factors, such as land use, soil type, and human activities should be evaluated to improve the accuracy of the models. Additionally, the applicability of the proposed solution in other geographic areas with different climate and topographic conditions should also be investigated. Finally, it is recommended that the algorithms used in this study be applied to other zoonotic diseases as well.

References

Adler B, 2015. History of leptospirosis and leptospira. In: Adler B, ed. Leptospira and Leptospirosis.Springer, Berlin, Heidelberg.

Ahmad A, Hassan M, Abdullah M, Rahman H, Hussin F, Abdullah H, Saidur R, 2014. A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renew Sustain Energy 33:102-9.

Ali M, Emch M, Donnay J-P, Yunus M, Sack R, 2002. Identifying environmental risk factors for endemic cholera: a raster GIS approach. Health & Place 8:201-10.

Anderson JA, 1995. An introduction to neural networks. MIT Press, Cambridge, MA.

Anselin L, 1995. Local indicators of spatial association - LISA. Geograph Anal 27:93-115.

Baggenstos R, Dahinden T, Torgerson PR, Bar H, Rapsch C, Knubben-Schweizer G, 2016. Validation of an interactive map assessing the potential spread of *Galba truncatula* as intermediate host of *Fasciola hepatica* in Switzerland. Geospat Health 11:137-43.

Barcellos C, Sabroza PC, 2001. The place behind the case: leptospirosis risks and associated environmental conditions in a flood-related outbreak in Rio de Janeiro. Cadern Saúde Pública 17:S59-67.

Ben-Hur A, Weston J, 2010. A user's guide to support vector machines. Data Mining Techn Life Sci 223-39.

Benesty J, Chen J, Huang Y, Cohen I, 2009. Noise reduction in speech processing. Springer Science & Business Media, Heidelberg.

Bradley AP, 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 30:1145-59.

Brown WM, Gedeon TD, Groves DI, 2003. Use of noise to augment training data: a neural network method of mineral–potential mapping in regions of limited known deposit examples. Nat Resource Res 12:141-52.

Burges CJ, 1998. A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discovery 2:121-67.

Caldas De Castro M, Singer BH, 2006. Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. Geograph Anal 38:180-208.

Carletta J, 1996. Assessing agreement on classification tasks: the kappa statistic. Computation Linguistic 22:249-54.

Duan K-B, Keerthi SS, 2005. Which is the best multiclass SVM method? An empirical study. International Workshop on Multiple Classifier Systems. Springer, Berlin, Heidelberg. pp 278-285.

Gracie R, Barcellos C, Magalhães M, Souza-Santos R, Barrocas PRG, 2014. Geographical scale effects on the analysis of leptospirosis determinants. Int J Environ Res Public Health 11:10366-83.

Hagan MT, Demuth HB, Beale MH, De Jesús O, 1996. Neural network design. PWS Publishing Company, Boston, USA.

Hippert HS, Pedreira CE, Souza RC, 2001. Neural networks for short-term load forecasting: A review and evaluation. IEEE Transactions Power Syst 16:44-55.

Honarmand H, Eshraghi S, Khorramizadeh M, Hartskeerl R, Ghanaei F, Abdolahpour G, Eshraghian M, 2007. Distribution of human leptospirosis in Guilan Province, Northern Iran. Iran J Public Health 36:68-72.

Hornik K, Stinchcombe M, White H, 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2:359-66.

Hsu C-W, Chang C-C, Lin C-J, 2003. A practical guide to SVM classification. Tech Rep Department of Computer Science and Information Technology, National Taiwan University. Available from: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Karande S, Bhatt M, Kelkar A, Kulkarni M, De A, Varaiya A, 2003. An observational study to detect leptospirosis in Mumbai, India, 2000. Archiv Dis Childhood 88:1070-5.

Lau C, Smythe L, Weinstein P, 2010. Leptospirosis: an emerging disease in travellers. Travel Med Infect Dis 8:33-9.

Lau CL, Clements AC, Skelly C, Dobson AJ, Smythe LD, Weinstein P, 2012. Leptospirosis in American Samoa–estimating and mapping risk using environmental data. PLoS Negl Trop Dis 6:e1669.

Mohammadinia A, Alimohammadi A, Habibi R, 2015. Assessment of environmental factors associated with rural endemics of Leptospirosis in Guilan Province, Iran. Proceedings of the International Conference on Research in Science and Technology. Kualalumpur, Malaysia; 14 December 2015. Available from: https://www.researchgate.net/publication/301204636_Assessment_of_environmental_factors_associated_with_rural_endemics_of_Leptospirosis_in_Guilan_Province_Iran

Mollalo A, Alimohammadi A, Shirzadi M, Malek M, 2015. Geographic information system-based analysis of the spatial and spatio-temporal distribution of zoonotic cutaneous leishmaniasis in Golestan Province, north-east of Iran. Zoonos Public Health 62:18-28.

Moore DA, Carpenter TE, 1999. Spatial analytical methods and geographic information systems: use in health research and epidemiology. Epidemiol Revi 21:143-61.

Moran PA, 1948. The interpretation of statistical maps. J Royal Statist Soc Ser B (Methodol) 10:243-51.

Pezeshki Z, Tafazzoli-Shadpour M, Nejadgholi I, Mansourian A, Rahbar M, 2016. Model of cholera forecasting using artificial neural network in Chabahar City, Iran. Int J Enteric Pathogen 4:23-30.

Raei M, Schmid VJ, Mahaki B, 2018. Bivariate spatiotemporal disease mapping of cancer of the breast and cervix uteri among Iranian women. Geospat Health 13:164-71.

Rafiei A, Zadeh-Omran AH, Babamahmoodi F, Navaei RA, Valadan R, Sari I, 2012. Review of leptospirosis in Iran. J Mazandaran Univ Med Sci 22.

Rajabi M, Mansourian A, Bazmani A, 2012. Susceptibility mapping of visceral leishmaniasis based on fuzzy modelling and group decision-making methods. Geospat Health 7:37-50.

Rajabi M, Mansourian A, Pilesjö P, Bazmani A, 2014. Environmental modelling of visceral leishmaniasis by susceptibility-mapping using neural networks: A case study in northwestern Iran. Geospat Health 9:179-91.

Rajabi M, Pilesjö P, Shirzadi MR, Fadaei R, Mansourian A, 2016. A spatially explicit agent-based modeling approach for the spread of Cutaneous Leishmaniasis disease in central Iran, Isfahan. Environ Model Software 82:330-46.

Ramezankhani R, Hosseini A, Sajjadi N, Khoshabi M, Ramezankhani A, 2017a. Environmental risk factors for the incidence of cutaneous leishmaniasis in an endemic area of Iran: A GIS-based approach. Spatial Spatio-temporal Epidemiol 21:57-66.

Ramezankhani R, Sajjadi N, Esmaeilzadeh RN, Jozi SA, Shirzadi MR, 2017b. Spatial analysis of cutaneous leishmaniasis in an endemic area of Iran based on environmental factors. Geospat Health 12:282-93.

Sousa S, Caeiro S, Painho M, 2002. Assessment of map similarity of categorical maps using Kappa statistics. ISEGI, Lisbon, Portugal.

Stevens KB, Gilbert M, Pfeiffer DU, 2013. Modeling habitat suitability for occurrence of highly pathogenic avian influenza virus H5N1 in domestic poultry in Asia: a spatial multicriteria decision analysis approach. Spatial Spatio-temporal Epidemiol 4:1-14.

Terpstra W, 2003. Human leptospirosis: guidance for diagnosis, surveillance and control. World Health Organization, Geneva, Switzerland. Available from: http://www.who.int/iris/handle/10665/42667

Vapnik VN, 1998. Statistical learning theory. Wiley, New York.

Vega-Corredor MC, Opadeyi J, 2014. Hydrology and public health: linking human leptospirosis and local hydrological dynamics in Trinidad, West Indies. Earth Perspect 1:14.

Williams D, Hinton G, 1986. Learning representations by back-propagating errors. Nature 323:533-6.