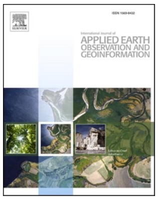
International Journal of Applied Earth Observations and Geoinformation 112 (2022) 102854

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Integration of machine learning algorithms and GIS-based approaches to cutaneous leishmaniasis prevalence risk mapping

Negar Shabanpour [a], Seyed Vahid Razavi-Termeh [a], Abolghasem Sadeghi-Niaraki [b,*], Soo-Mi Choi [b], Tamer Abuhmed [c]

[a] Geoinformation Tech. Center of Excellence, Faculty of Geodesy and Geomatics Engineering, K.N. Toosi University of Technology, Tehran 19697, Iran
[b] Department of Computer Science and Engineering & Convergence Engineering for Intelligent Drone, Sejong University, Seoul 143-747, Republic of Korea
[c] College of Computing and Informatics, Sungkyunkwan University, Suwon 16419, Republic of Korea

ARTICLE INFO

Keywords:
Cutaneous leishmaniasis
Machine learning
Spatial modelling
Environmental factors
Health geography

ABSTRACT

Cutaneous leishmaniasis is a complex infection that is caused by different species of Leishmania and affects more than 2 million people in 88 countries. Identifying the environmental factors affecting the occurrence of cutaneous leishmaniasis and preparing a risk map is one of the basic tools to control and manage this disease. The aim of this study was a spatial prediction of cutaneous leishmaniasis in Isfahan province, Iran using three machine learning algorithms (decision tree (DT), support vector regression (SVR), and linear regression (LR)). The spatial database was created using data collected on the number of diseases in Isfahan province from 2011 to 2018, as well as ten environmental parameters (temperature, humidity, rainfall, altitude, slope, wind speed, normalized difference vegetation index (NDVI), number of sunny days, number of frosty days, and distance to stream) that affect the incidence of leishmaniasis. Furthermore, the fuzzy method was employed in this study to reduce uncertainty and evaluate the effect of environmental factors on disease prevalence. Using the holdout method and 70:30 ratios, the data were used to model and prepare a leishmaniasis prediction map and evaluate the results, respectively. The accuracy of the maps satisfied with the DT, SVR, and LR algorithms was 0.951, 0.934, and 0.914, respectively, according to the receiver operating characteristic (ROC) curve and area under the curve (AUC). Furthermore, the eastern and southern parts of the province have the lowest risk of leishmaniasis. The result of this issue is the identification of high-risk areas of the disease and increase life and peace for people in the community.

1. Introduction

The world health organization (WHO) considers leishmaniasis to be one of the world's six major tropical diseases (WHO, 2019). Leishmaniasis parasites cause the disease, which is transferred to humans and animals by sandflies of the phlebotomine subfamily (Länger et al., 2012). The annual leishmaniasis incidence in the world is expected to be between 700,000 and 1.2 million people, according to a WHO report published in 2019 (Mohammadbeigi et al., 2020). Leishmaniasis is the name of a group of parasitic diseases, which appear in various forms owning to the host's immune response, including visceral leishmaniasis (VL) and cutaneous leishmaniasis (CL) (Hartley et al., 2014). VL, also known as visceral, black fever, or cattle fever, is the most severe form of this disease. This type of disease is relatively rare and accounts for only 10% of all leishmaniasis worldwide. CL is the most frequent type of leishmaniasis, which is seen in many subtropical areas. Since 1930, 50 species of sandflies have been reported to cause CL in Iran (Sofizadeh et al., 2017). In Iran, the leishmaniasis vector is species of the Phlebotomus papatasi sand-fly. In 17 of Iran's 31 provinces, this disease is endemic (Khamesipour et al., 2020). The majority of leishmaniasis infections are found in Iran's eastern, central, and southern regions. 62.6% of cases of leishmaniasis occur in rural areas (Mohammadbeigi et al., 2020). Isfahan, in central Iran, has long been regarded as one of the most dangerous leishmaniasis hotspots (Ramezankhani et al., 2017). Despite extensive preventive steps taken by health officials and ongoing research efforts over the last 30 years, the incidence of cases in Isfahan remains high (Rajabi et al., 2016). People who have visited or lived in disease-endemic areas have the greatest leishmaniasis rate, with 7.2% of leishmaniasis cases coming from non-endemic parts of the Isfahan province (Nilforoushzadeh et al., 2015). Economic, social, environmental, rapid

---

[*] Corresponding author.
E-mail addresses: a.sadeghi@sejong.ac.kr (A. Sadeghi-Niaraki), smchoi@sejong.ac.kr (S.-M. Choi), tamer@skku.edu (T. Abuhmed).

https://doi.org/10.1016/j.jag.2022.102854
Received 7 December 2021; Received in revised form 15 May 2022; Accepted 30 May 2022

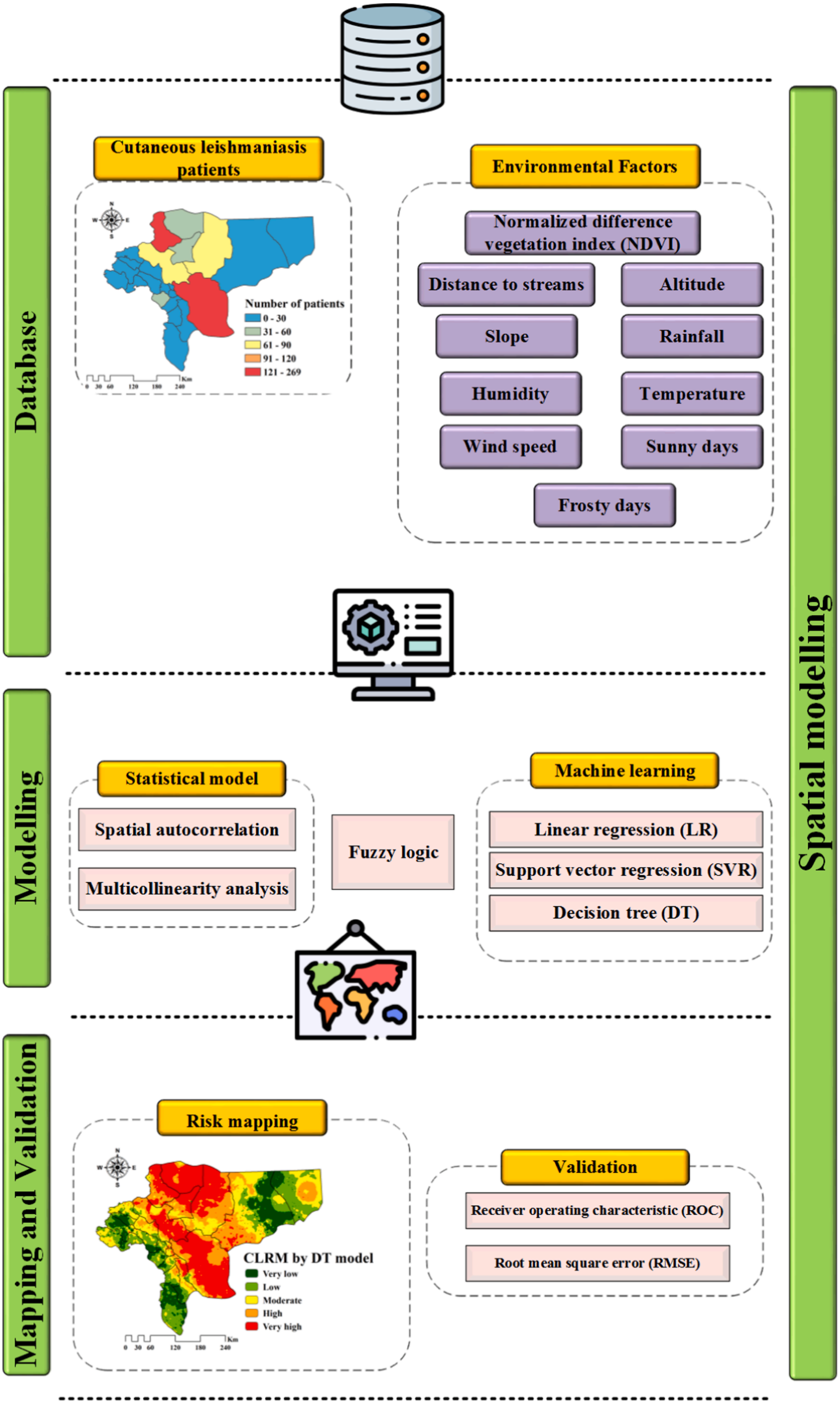


Fig. 1. Research methodology.

Fig. 2. Location of Isfahan province in Iran.

urban growth, and ecological factors all influence the prevalence of diseases like leishmaniasis. Among these factors, climatic elements are a critical factor in the incidence of leishmaniasis (de Santana Martins Rodgers et al., 2019).

In endemic locations, climatic patterns can create ideal circumstances for the spread and transmission of CL. Rainfall and humidity have an impact on the epidemiology of sandflies because they affect population density, distribution, and survival (Mohammadbeigi et al., 2020).

Geographic information system (GIS), owing to its useful applications in recording and storing, analysing, managing, and displaying prevalence of disease information, could be used to model the spatial distribution of diseases, Spatio-temporal relationships, and prepare prediction maps (Ramezankhani et al., 2017). Several studies have employed spatial modelling and analysis to study the impact of environmental factors on leishmaniasis spread so far. Moradiasl et al. (2018) examined the relationship between environmental variables and the leishmaniasis incidence in Ardabil province, Iran. Tabasi et al. (2020) used spatial modelling of leishmaniasis in Golestan province, Iran. Gherbi et al. (2020) studied the spatial distribution of leishmaniasis sand-flies from 2016 to 2017 in North-Eastern Algeria. In Bahia, Brazil, Adegboye and Adegboye (2017) used the ecological niche model (ENM) to predict the distribution of VL vectors and identify the factors impacting disease prevalence. Iliopoulou et al. (2018) studied leishmaniasis in Isfahan province, Iran and examined the relationship between the disease prevalence, age, and gender. Melo et al., (2018) investigated the distribution pattern of VL sandflies in Athens, Greece, and the relationship between socioeconomic parameters. Using a geographically weighted regression and least squares regression, Mollalo et al. (2018) evaluated the relationship between leishmaniasis cases and environmental factors in Isfahan Province, Iran. Wu et al. (2019) examined the relationship between vegetation and the prevalence of leishmaniasis in southern Brazil. In Golestan province, Iran, Akhavan et al. (2014) utilized logistic regression, random forest (RF), and support vector machine (SVM) algorithms to model and forecast the presence and absence of sand-flies that cause leishmaniasis. So far, researches on the prevalence of leishmaniasis have focused on the relationship between the prevalence of the disease and environmental factors affecting it. To the best of the authors' knowledge, research to date has not examined the development of leishmaniasis risk prediction maps by GIS-based machine learning algorithms. Machine learning is a way of creating complicated algorithms and models for prediction in data analysis (Battineni et al., 2020). Machine learning algorithms are a good tool in the spatial modelling and prediction of many diseases (Rajabi et al., 2016). Machine learning algorithms in the study of diseases have received much attention in recent years. According to previous research, numerous methods have been used to investigate the effect of environmental parameters on leishmaniasis, but intelligent methods have been shown to provide more accurate results. In addition, each of the non-intelligent methods has its own strengths and weaknesses, which may work well in one study and not in another. Therefore, the desire to combine intelligent methods is growing to achieve a sustainable and desirable answer (Mohebali et al., 2020).

Because leishmaniasis is related to a variety of environmental parameters, it is difficult to estimate the effective parameters and the relationship between them in the prevalence of leishmaniasis because hidden information is difficult to identify in a large data set, and machine learning is a solution. Three main criteria are addressed when diagnosing and analyzing a disease: time, place, and person. Most reasons for creating disease are location-based, in other words, their distribution and focus on the different locations are different (Razavi-Termeh et al., 2022). Given that spatial and disease-related information contains a large amount of data, machine learning is an approach to extracting information and knowledge from this data. Integration of

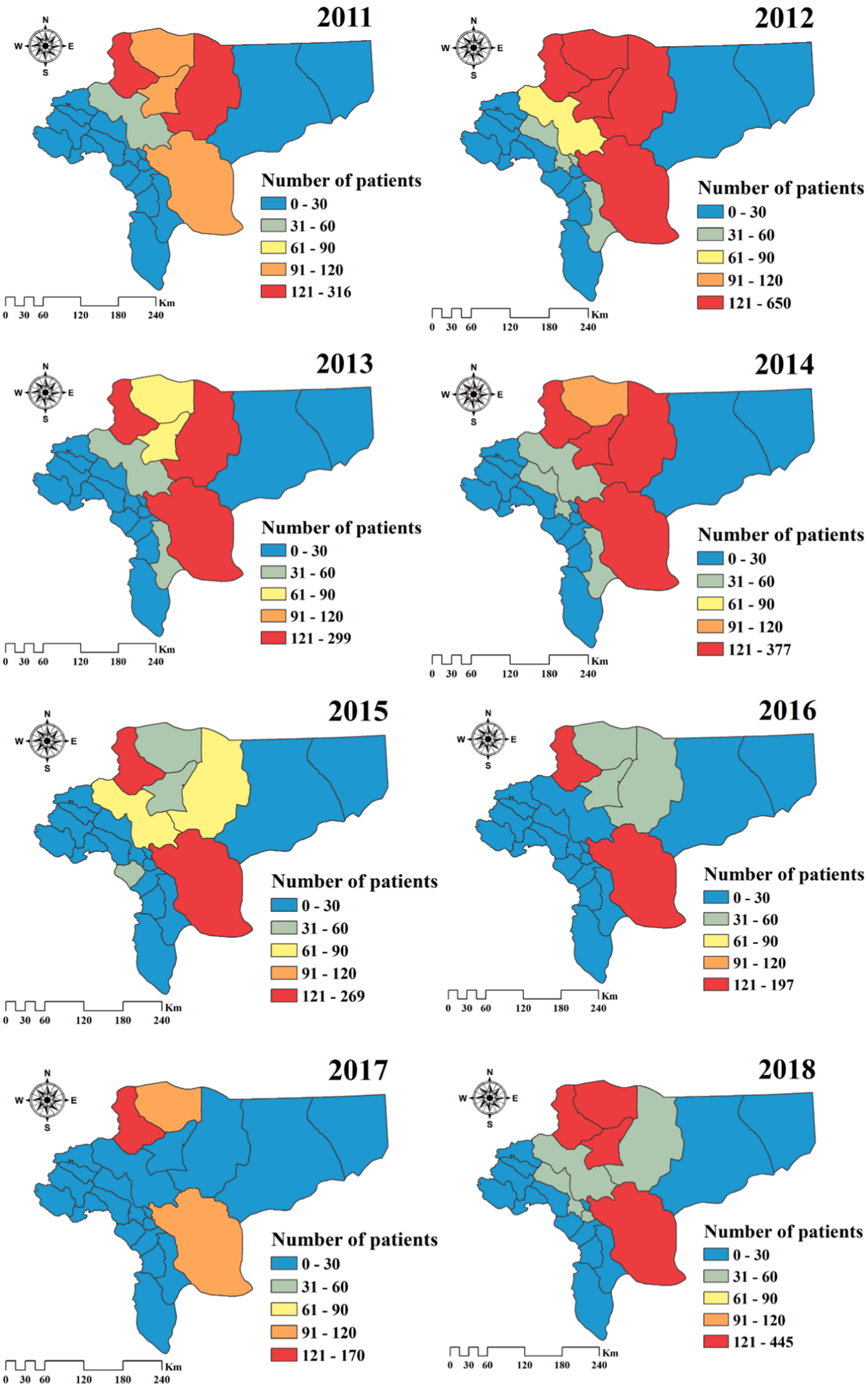


Fig. 3. The number of leishmaniasis patients.

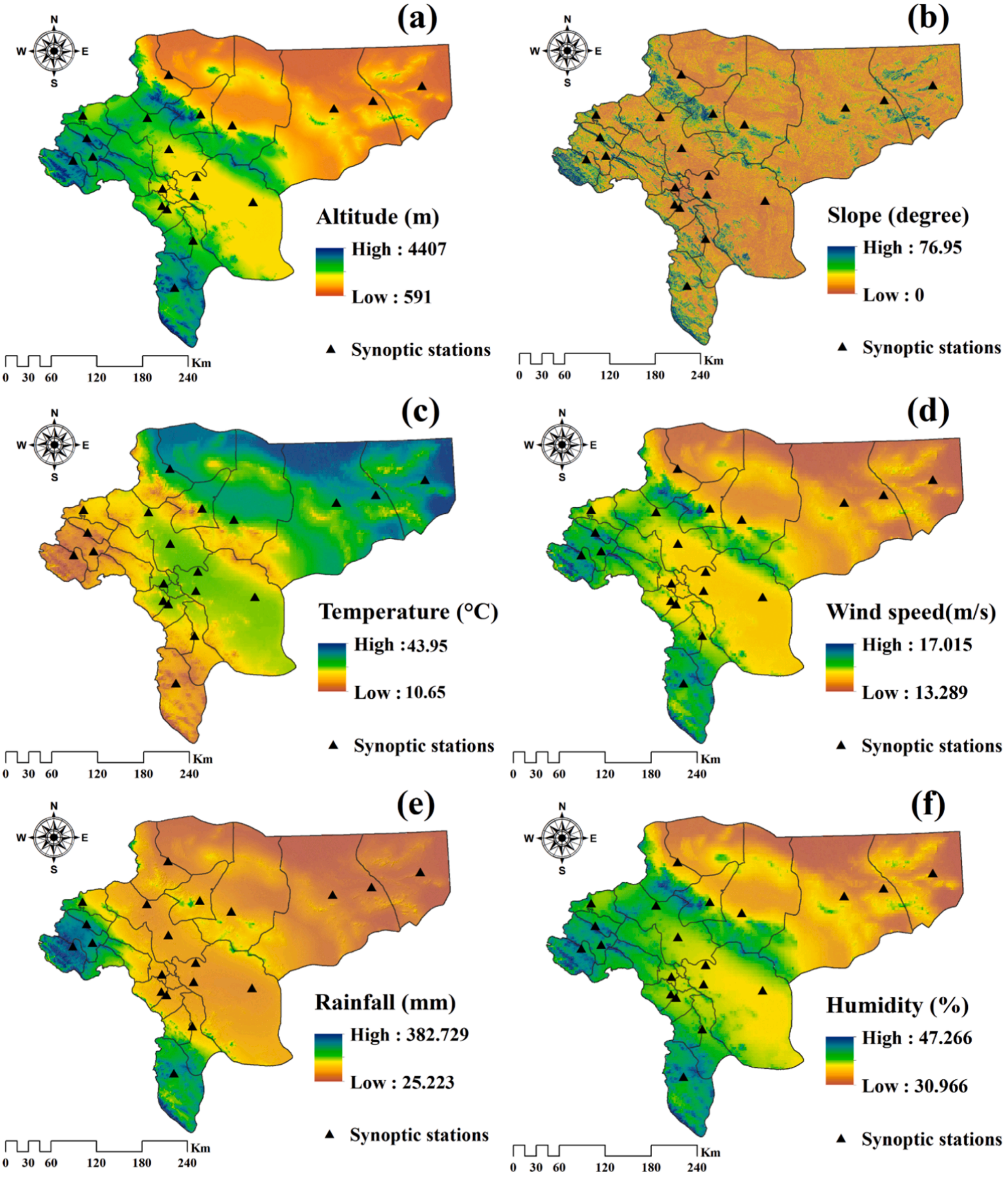


Fig. 4. Environmental factors affecting leishmaniasis: a) Altitude b) Slope c) Temperature d) Wind speed e) Rainfall f) Humidity g) Sunny days h) Frosty days i) NDVI, and j) Distance to stream.

machine learning algorithms and GIS makes this possible for health researchers to merge data for different resources and have done different analyses of spatial information (VoPham et al., 2018). The objectives of this paper were to use three machine learning algorithms (decision tree (DT), support vector regression (SVR), and linear regression (LR) to investigate the effect of environmental factors on the prevalence of leishmaniasis, spatial autocorrelation, spatial modelling, and the preparation of a leishmaniasis risk map in Iran's Isfahan province. The innovation of the present study was in using three machine learning algorithms to prepare a map to predict the risk of leishmaniasis and Spatio-temporal analysis of the disease in the Isfahan province of Iran.

2. Material and methods

2.1. Methodology

This study was carried out in four steps, as shown in Fig. 1. The first step was to develop a spatial database that included environmental parameters and the number of leishmaniasis cases. Spatial autocorrelations, multicollinearity analyses, and fuzzy method were done in the

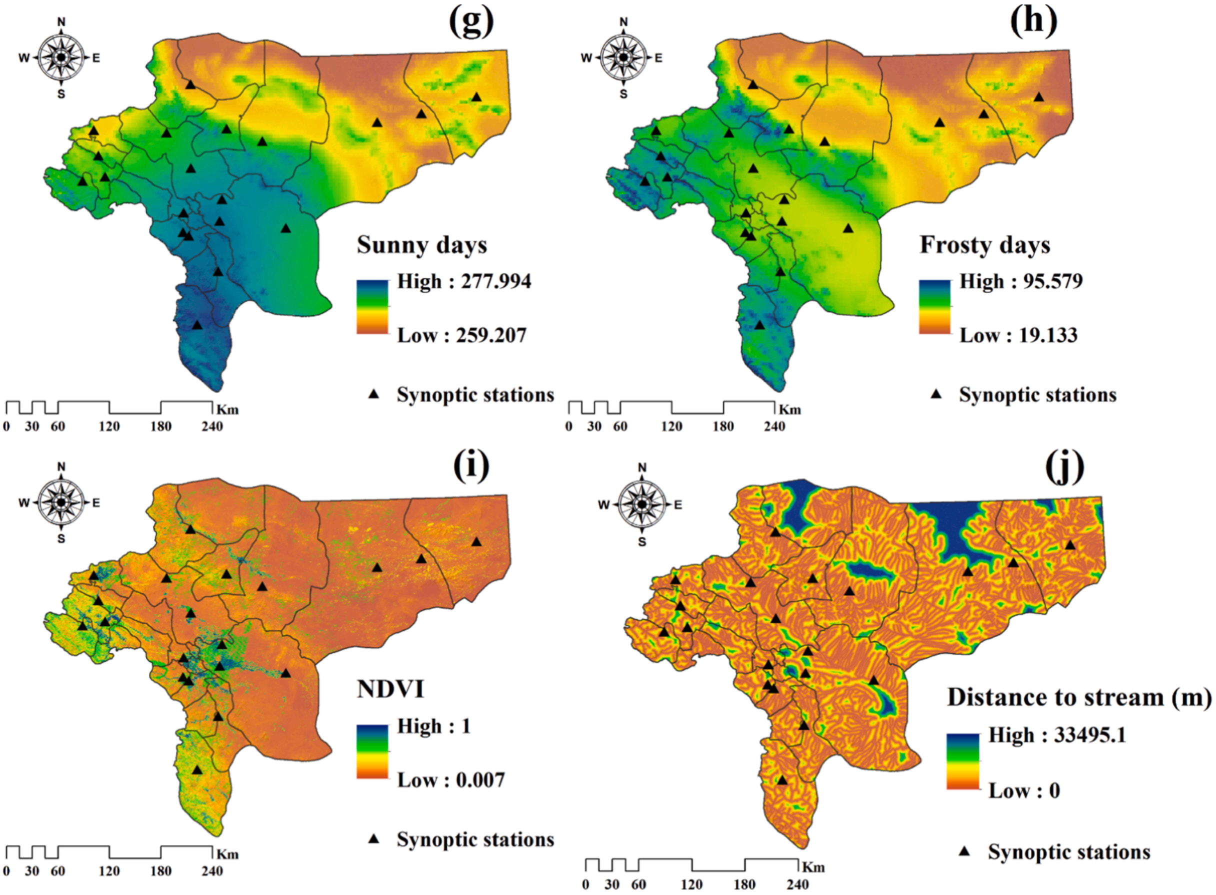


Fig. 4. (*continued*).

Table 1

Result of multicollinearity between independent factors.

Factors	Tolerance	VIF
NDVI	0.883	1.132
Distance to stream	0.81	1.234
Slope	0.729	1.372
Humidity	0.492	1.821
Wind speed	0.431	2.21
Temperature	0.761	1.411
altitude	0.732	1.834
Rainfall	0.425	2.02
Frosty days	0.569	1.693
Sunny days	0.801	1.389

Table 2

Results of the type of disease distribution with Moran's I index.

Year	Index value	variance	Expected value	Z-score	P-value	Type of distribution
2011	0.173	0.006	−0.047	20.65	<0.05	Cluster
2012	0.227	0.006	−0.047	30.39	<0.05	Cluster
2013	0.131	0.007	−0.047	20.07	<0.05	Cluster
2014	0.148	0.007	−0.047	20.22	<0.05	Cluster
2015	0.044	0.006	−0.047	10.12	<0.05	Cluster
2016	0.1	0.006	−0.047	10.87	>0.05	Random
2017	0.095	0.006	−0.047	10.87	>0.05	Random
2018	0.2	0.007	−0.047	20.88	<0.05	Cluster

second step. Three machine learning algorithms (DT, SVR, and LR) were used in the third step to perform spatial modeling and prepare a leishmaniasis prediction map. The modeling results were evaluated using the root mean square error (RMSE) and receiver operating characteristic (ROC) indices in the fourth step.

2.2. Study area

Isfahan province is in the center of Iran, with an average area of 106786 km^2 with latitudes ranging from 30° 43′ to 34° 27′ north and longitudes ranging from 49° 36′ to 55° 31′ east. The range of altitude in the study areas is between 591 and 4407 m. The average annual temperature is 16.7 °C, with a maximum temperature of 40.6 °C and a minimum temperature of −10.6 °C. The province has 76 frosty days each year, with an average annual rainfall of 116.9 mm. The study area is shown in Fig. 2.

The west of Isfahan province is mountainous and has several elevations, long winters, more rainfall, and a low density of Zagros forests. The breeze is moderate and gentle in these areas, and the average annual temperature is mostly 12 °C.

2.3. Spatial database

2.3.1. Leishmaniasis data

In this research, all recorded information from patients with leishmaniasis from 2011 to 2018 in the counties of Isfahan province was obtained from the Infectious Diseases Management Center (IDMC). According to statistics in each county, the number of diseases increased from 2011 to 2013, then decreased from 2014 to 2017. In 2018, the number of diseases in the study area increased (Fig. 3). Towns with a large number of diseases and higher risk were considered as model training points.

2.3.2. Environmental factors

In this study, to model the spread of leishmaniasis, some parameters

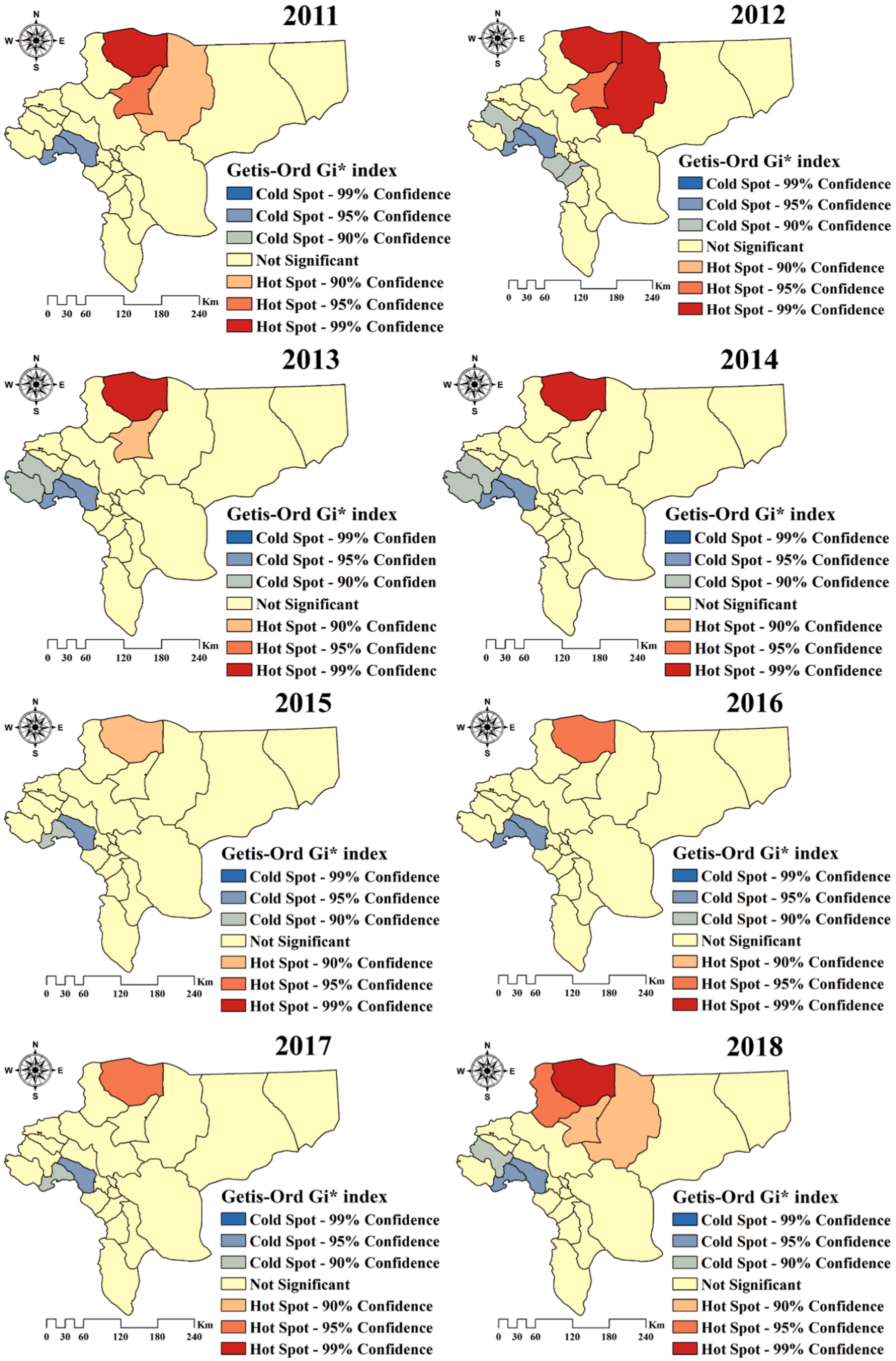


Fig. 5. Hot spot analysis of disease with Getis-ord Gi * index.

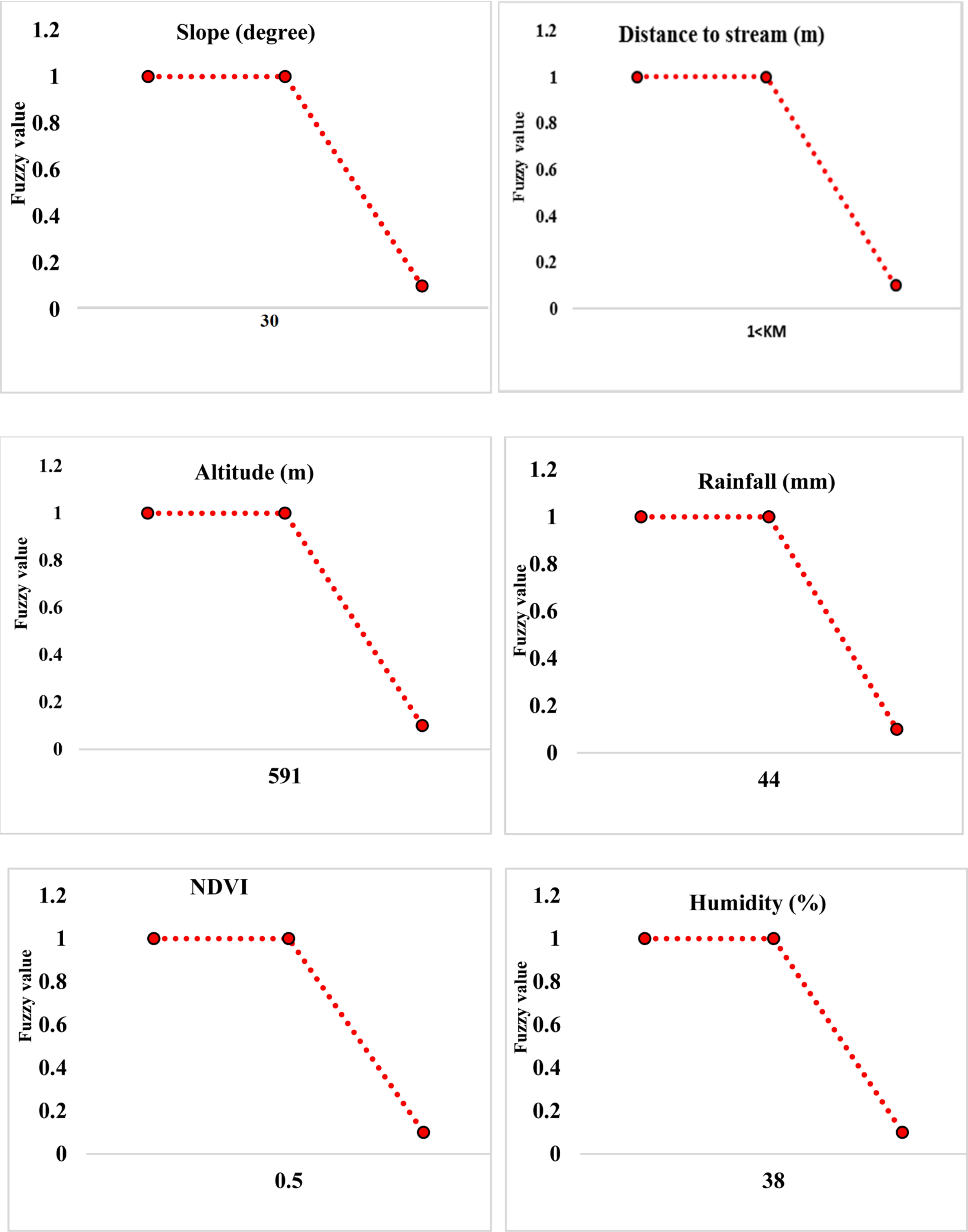


Fig. 6. Fuzzy function used in fuzzification of the effective factors.

such as temperature, wind speed, rainfall, humidity, number of frosty and sunny days, distance to stream, altitude, normalized difference vegetation index (NDVI), and slope were considered (Akhavan et al. 2014, Mollalo et al. 2018). Climatic factors of temperature, wind speed, rainfall, humidity, and the number of frosty and sunny days were obtained from 21 synoptic stations in Isfahan province on an annual average from 2011 to 2018. A 30*30 m pixel size map of these factors was created using the cokriging interpolation method in ArcGIS 10.3 software. For cokriging analysis, synoptic station data and the altitude of the study area were used.The digital elevation model (DEM) for the study area was built using ASTER images with 30-m pixel size and then used to calculate altitude and slope. The NDVI factor was calculated

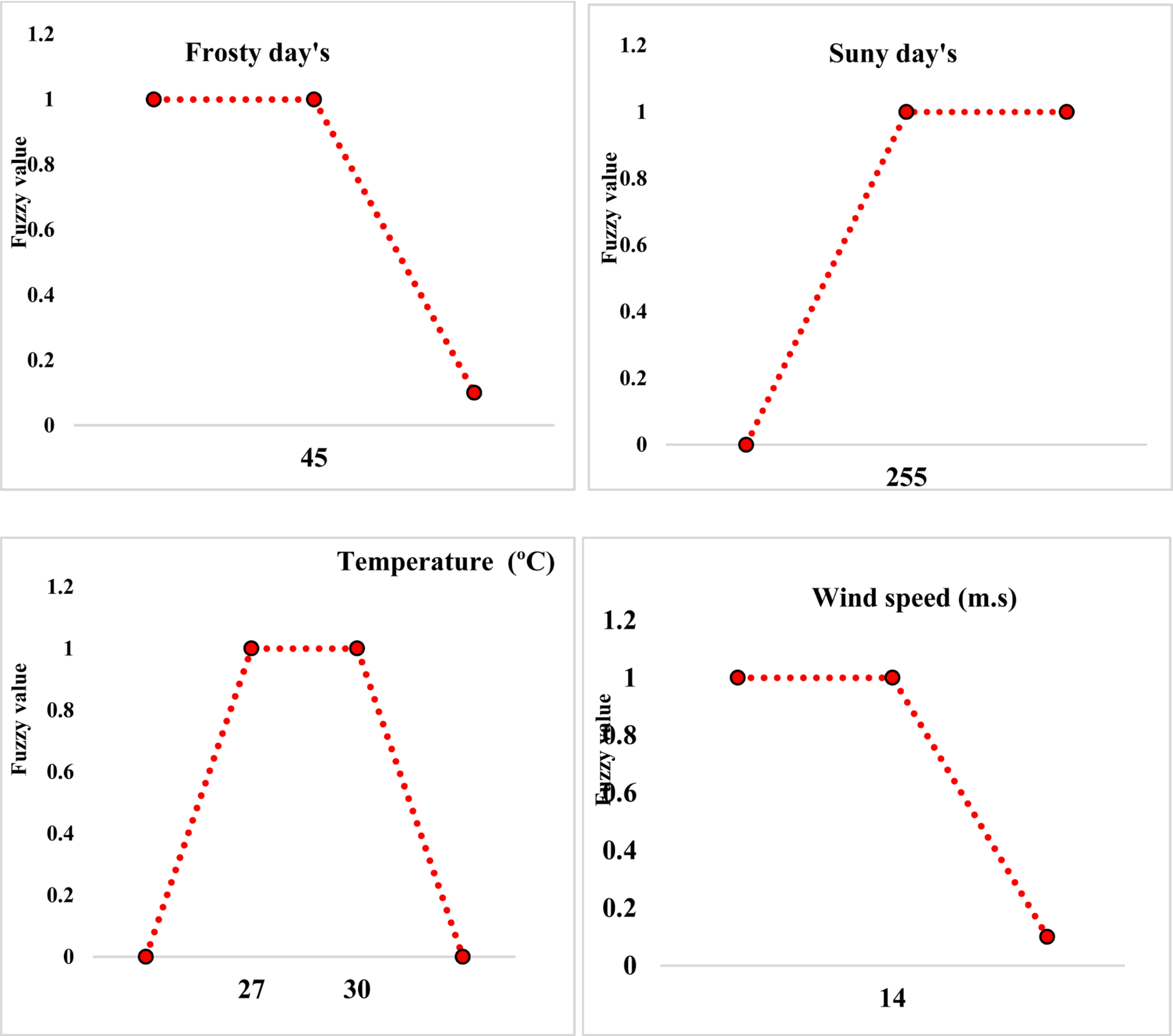


Fig. 6. (*continued*).

using Landsat 8 images in the Google Earth Engine (GEE) platform with a pixel size of 30 m. The stream layer was prepared using a 1: 100,000 topographic layer. The distance from the stream was then calculated using ArcGIS 10.3′s Euclidean distance tool. Environmental factors are shown in Fig. 4.

2.4. Methods

2.4.1. Multicollinearity analysis

The variance inflation factor (VIF) is an index used to indicate the presence or absence of a collinearity relationship between independent variables. This index expresses how much of the change in the coefficients met for each independent variable has increased to establish a collinearity relationship between the independent variables (Razavi-Termeh et al., 2020). In other words, when a multiple regression algorithm has multicollinearity, the resulting model coefficients are invalid (Marcoulides and Raykov, 2019). A collinearity between the independent variables exists when the VIF is greater than 5 and the tolerance value is less than 0.1 (Waitz et al., 2019). Multicollinearity analysis were performed in SPSS software.

2.4.2. Geostatistical analysis

Spatial autocorrelation is the correlation between the values of a variable that is severely attributed to their near spatial locations on a two-dimensional level (Mondini et al., 2017). This global statistical analysis describes the overall distribution of a variable to determine whether spatial clusters exist in a larger area and uses a single value to reflect the degree of autocorrelation of the area. In this study, Moran's I and Getis-ord Gi* indices were used to investigate spatial autocorrelation and hot spot analysis of leishmaniasis. The Moran's I index was employed to look at the kind of disease distribution (cluster or random) between 2011 and 2018. The Getis-ord Gi* analysis was also used to determine the disease's high-risk areas and how the disease spread over time. The purpose of these two indicators is to determine the type of distribution and clusters of the disease and how the disease spreads during this period (2011 to 2018). Spatial autocorrelation analyses were performed in ArcGIS 10.3 software.

- Moran's I index

Moran's I index is a statistical indicator for studying spatial autocorrelation (Mondini et al., 2017). The Moran index measures autocorrelation based on location and property values to determine whether the pattern is clustered, dispersed, or random (Sharma et al., 2018). Moran's spatial autocorrelation tool is an inferential statistic, which means that the analysis results are always interpreted within the framework of the null hypothesis. Spatial autocorrelation exists when there are systematic spatial variations in variable values. This change

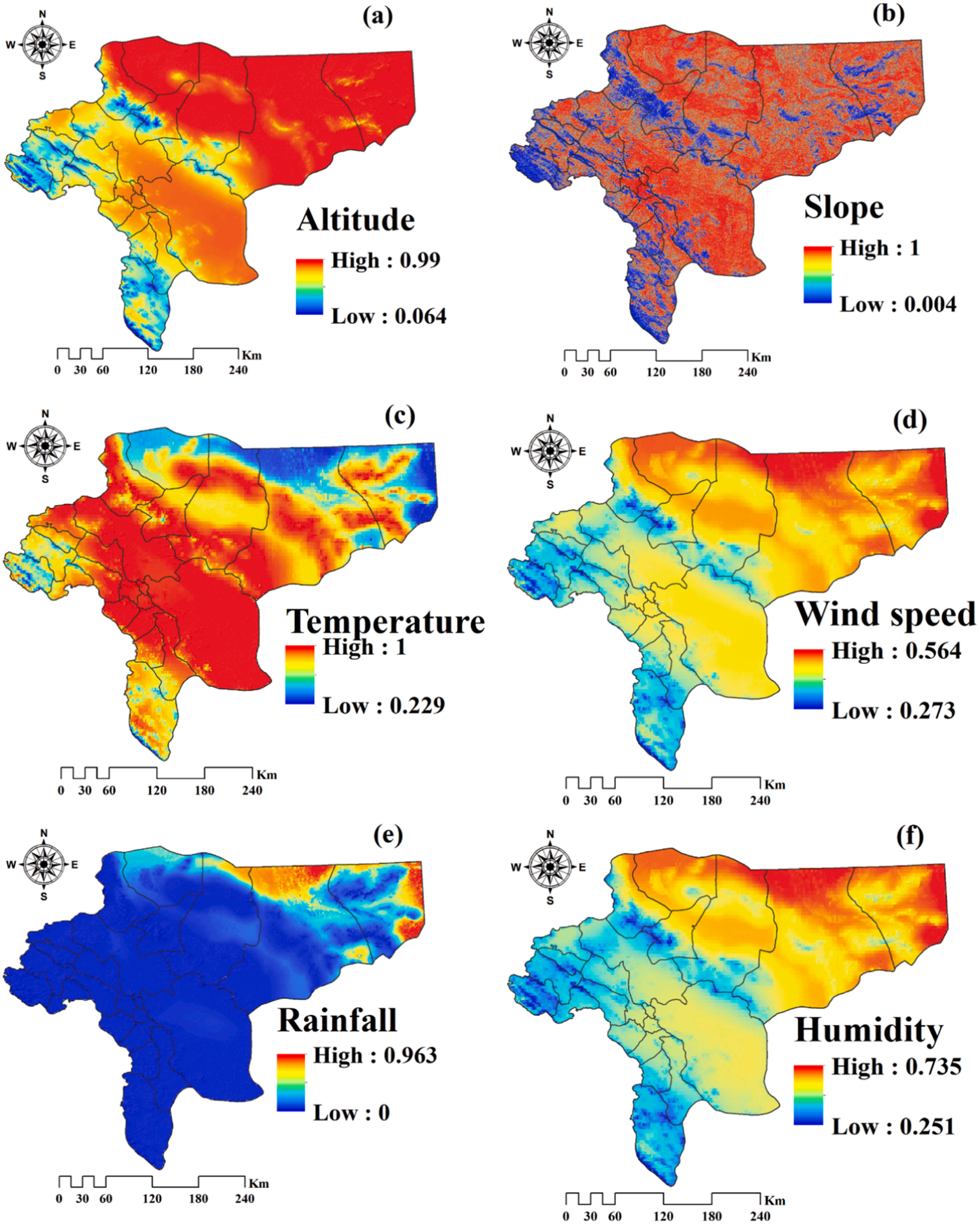


Fig. 7. Fazzification of environmental factors: a) Altitude b) Slope c) Temperature d) Wind speed e) Rainfall f) Humidity g) Sunny days h) Frosty days i) NDVI, and j) Distance to stream.

can be positive or negative spatial autocorrelation. Positive spatial autocorrelation occurs when adjacent neighborhood situations have comparable characteristic values, and negative spatial autocorrelation occurs when there is a distinct difference between neighborhood values. When there is no significant spatial autocorrelation, a random spatial distribution pattern is considered. Global spatial autocorrelation indices have been used to assess their similarity (Razavi-Termeh et al., 2021). For Global Moran’s I statistics, the null hypothesis states that the property being analyzed is distributed randomly among the properties in the study area (Mathur, 2015). The Moran index is calculated from Eq. 1:

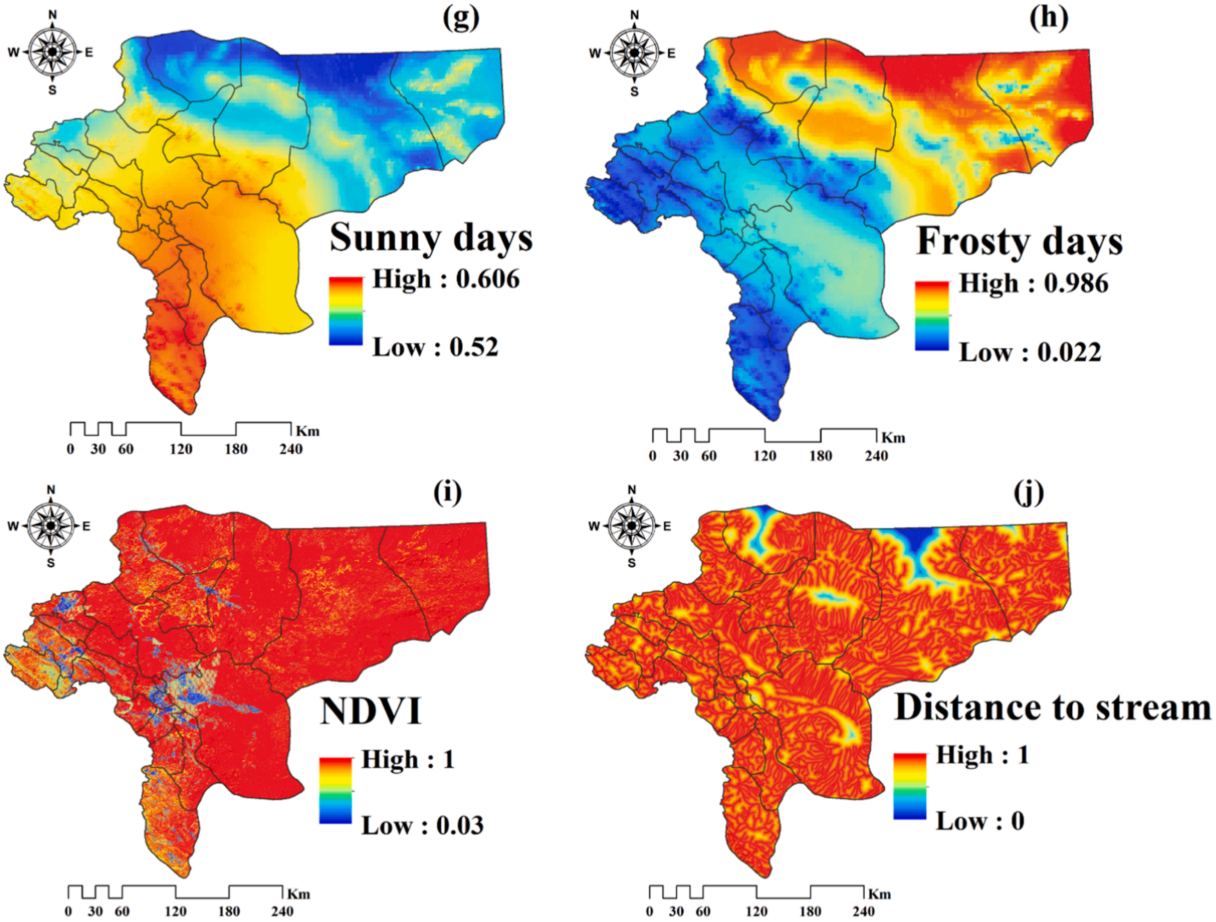


Fig. 7. (*continued*).

Table 3
Results of modeling accuracy with RMSE index.

Model	Training	Validation
DT	0.228	0.289
SVR	0.274	0.324
LR	0.292	0.347

$$I = \frac{N\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\right)\sum_{j=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{1}$$

Where N is the number of observations of the whole region, $\bar{x}$ is the mean of the variable, x_i is the variable value at a particular location, x_j is the variable value at another place, and W_{ij} is a spatial weight between locations of I relative to j (Khashoggi and Murad, 2020).

- Getis-ord Gi* index

The local Getis-ord Gi* index is used to investigate the degree of autocorrelation and spatial differences between adjacent cell values in a geographical area. This index is used to find "hot areas" (positive autocorrelation) and "cold spots" (negative autocorrelation) (Mathur, 2015). In this index, positive Z-score values indicate hot spots, and negative Z-score values indicate cold spots. The Z-score is a standardized score that illustrates the degree of disease concentration or distribution. The farther this value is from zero, the more indicative that the disease phenomenon is clustered or dispersed. If the Z sign is positive, it means that there is clustering among the areas where the disease has a high value, and if it is negative, it means that there is dispersing among the areas where the disease has a low value (Anselin, 1995). The Getis-ord Gi* index was calculated using Eq. 2:

$$G_i^* = \frac{\sum_{j=i}^{n} W_{i,j} x_j - \bar{x}\sum_{j=i}^{n} W_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)\right]}{n-1}}} \tag{2}$$

2.4.3. Fuzzy logic method

Fuzzy theory is a method that incorporates individual and ambiguous judgments about a unique phenomenon into probabilistic or mathematical models and provides the basis for inference, control, and decision-making in uncertain conditions.Unlike classical logic, which has two values of zero and one, fuzzy logic shows its value as a percentage of membership in the range (1, 0) (Imeni, 2020). Uncertainty assessment is done by assigning a number between 0 and 1 to this member. If this number is equal to zero, it can be stated with certainty that the target member does not belong to the set, and if this number is equal to one, it can be claimed that the target member belongs to that set (Yariyan et al., 2020). It's essential to select the appropriate fuzzy function for standardizing criteria (Khorram et al., 2015). Standardization using fuzzy set membership indicates a special relationship between criteria and decision set (Jiang and Eastman, 2000). The fuzzy maps generated for each criterion have membership values between zero and one, with zero showing the area with the lowest probability of disease prevalence and one indicating the probability of higher disease prevalence. In this work, fuzzy maps, related to environmental parameters, were used as input for the modelling process. Therefore, using the fuzzy method for input values, the machine learning algorithm more

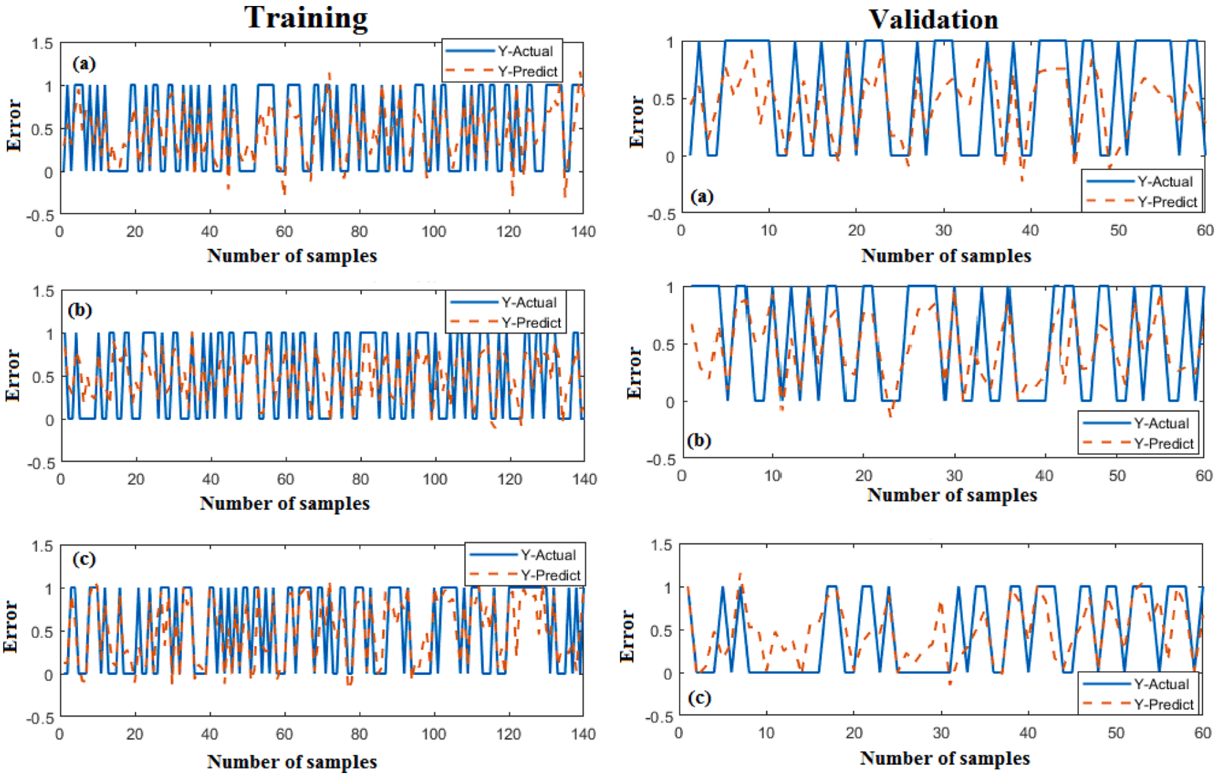


Fig. 8. Models predicting error by: a) DT, b) SVR, and c) LR.

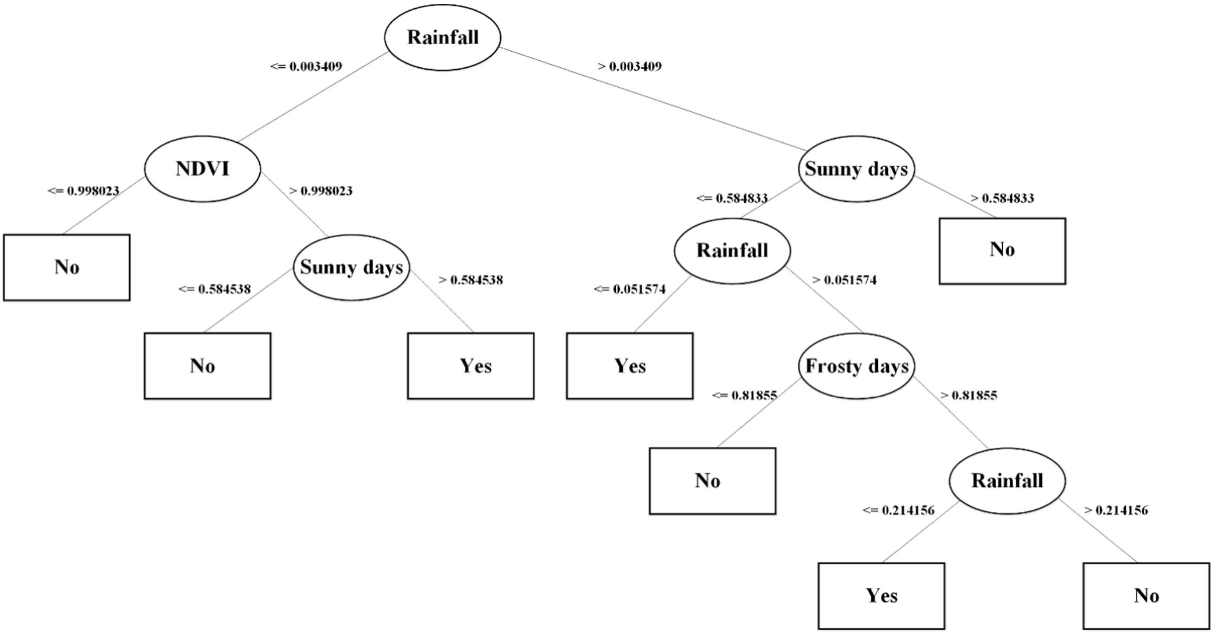


Fig. 9. An example of a tree in the DT algorithm.

accurately detects the relationship between the effective parameters and helps to improve the prediction accuracy. Fuzzification analysis were performed in ArcGIS 10.3 software.

2.4.4. Machine learning algorithms

Leishmaniasis risk map was done utilizing the Scikit -Learn library in Python and three machine learning algorithms: SVR, DT, and LR. The following describes each of these algorithms.

- Support vector regression (SVR) algorithm

This algorithm is based on statistical learning theory, which uses structural risk minimization and leads to an overall optimal response (Kohli and Arora, 2018). The aim of SVR algorithm is to discover the function f(x) for X training patterns so that it has the highest margin of training values Y. In other words, the SVR algorithm is a model that fits a ε thickness curve to the data so that the minimum error occurs in the training data. The SVR algorithm uses f(x) as Eq. 3 for predicting:

Table 4
Results of independent factor coefficients in LR algorithm.

	Unstandardized Coefficients		Standardized Coefficients	t	Sig0.
	B	Std0. Error	Beta		
(Constant)	−3.747	0.891		−0.709	0.479
NDVI	0.211	0.141	0.093	1.504	0.134
Distance to stream	−0.132	0.719	−0.12	−0.184	0.854
Slope	0.081	0.087	0.063	0.928	0.354
Humidity	−34.836	0.98	−0.871	−60.542	0.042
Wind speed	20.021	0.451	0.981	−2.050	0.153
Temperature	2.955	0.612	0.789	4.825	0.000
Altitude	−0.872	0.774	−0.321	−1.127	0.261
Rainfall	0.847	0.908	0.133	0.933	0.352
Frosty days	7.24	0.514	0.921	3.877	0.000
Sunny days	1.584	0.687	0.845	0.466	0.642

$$f(x) = WX + b \tag{3}$$

W is the weight of vector X, and b is the bias value. To minimize test error, the expression $\frac{1}{2}\|w\|^2$ should be minimized, which requires minimizing the norm of the weight vector. Vapnik (1999) used a new error function for the support vector machine in regression problems, which ignores errors at a certain distance from the actual values. The as mentioned deviation is defined as Eq. 3 and is considered to include deficiency variables ξ_i^+ and ξ_i^- in Eq. 4. Finally, based on the principle of structural error minimization, the error range is optimized by the optimization problem in Eq. 5 (Awad and Khanna, 2015).

$$|\boldsymbol{\xi}|_\varepsilon = \begin{cases} 0 if \quad |\xi| \le \varepsilon \\ |\xi| - \varepsilon otherwise \end{cases} \tag{3}$$

$$Minimize: \quad \frac{1}{2}(w.w) + C\sum_{i=1}^{N}(\xi_i^+ + \xi_i^-) \tag{4}$$

$$Constraints: \quad \begin{cases} W.X_i + b - y_i \le \varepsilon + \xi_i^+ i = 1.2.3.\cdots.N \\ y_i - (W.X_i + b) \le \varepsilon + \xi_i^- i = 1.2.3.\cdots.N \\ \xi_i^+ \ge 0.\xi_i^- \ge 0 i = 1.2.3.\cdots.N \end{cases} \tag{5}$$

W is the regression coefficient vector, ε is the margin deviation and C is the penalty parameter.

- Linear regression (LR) algorithm

In small sample sizes, regression algorithms based on least-squares fitting are employed to investigate the correlations between independent and dependent variables (Yashon et al., 2020). Since there may be many other unknown factors in addition to independent variables, which effectively determine the dependent variable, the regression algorithm was considered with the most appropriate number of independent variables and the error rate as a representative for other random factors that are unrecognizable (Alexopoulos, 2010).

- Decision tree (DT) algorithm

DT is a decision-based hierarchical method that splits independent variables into homogeneous sections. The goal of this method is to find a set of decision rules that can predict the outcome based on input data.

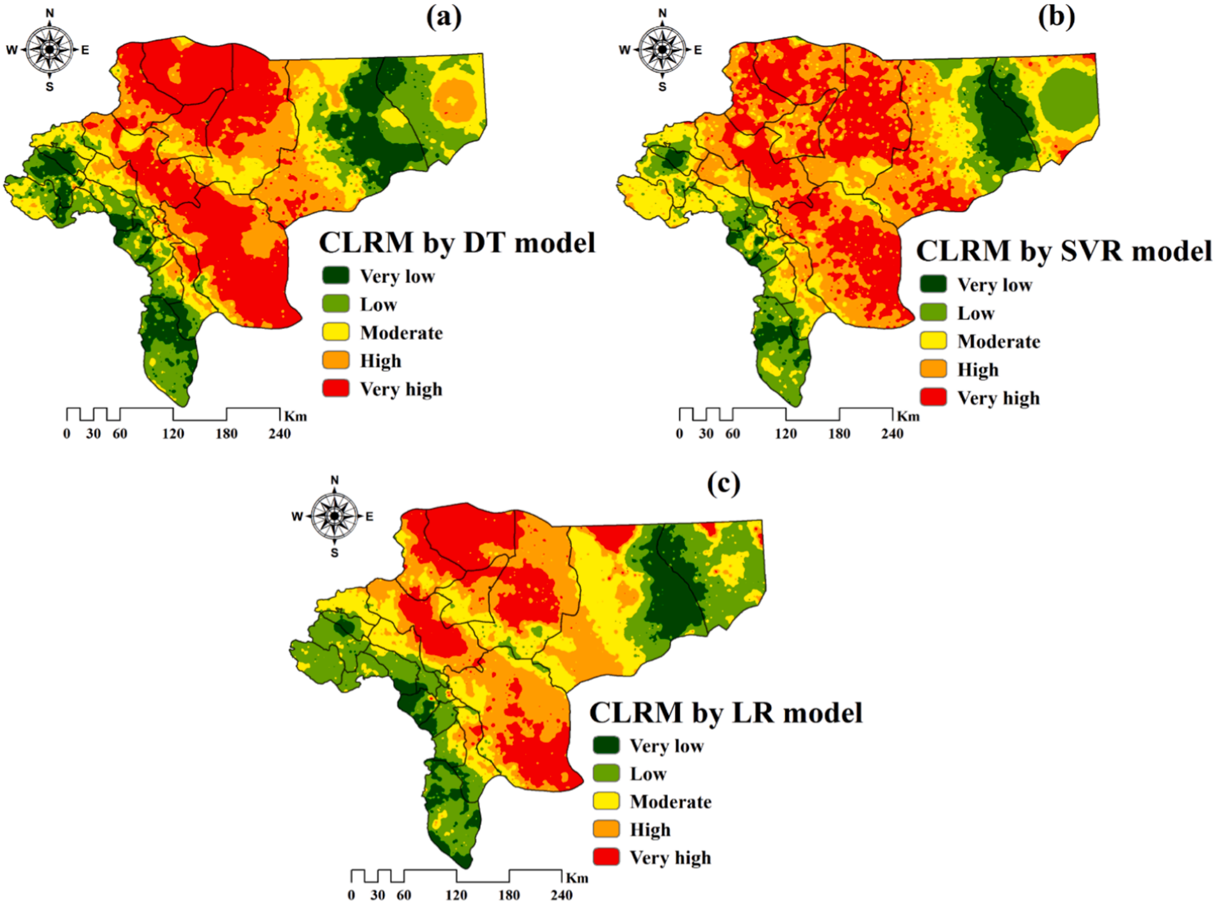


Fig. 10. Cutaneous leishmaniasis risk mapping by: a) DT, b) SVR, and c) LR algorithms.

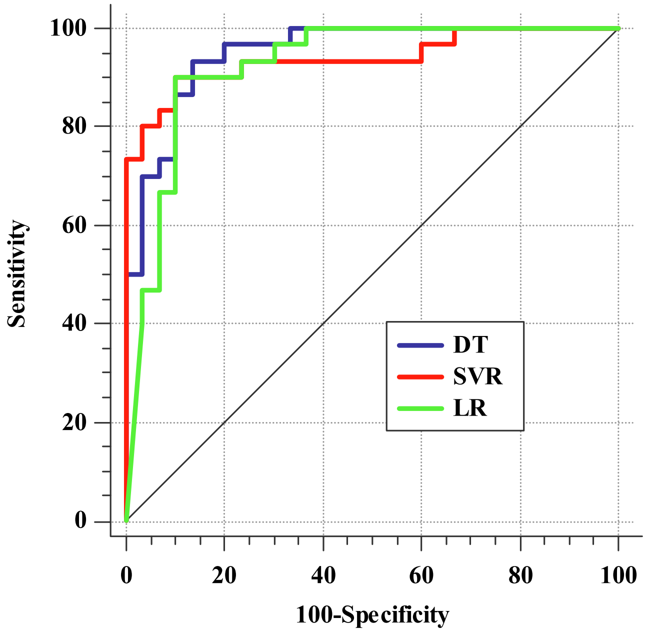


Fig. 11. Results of the accuracy of three algorithms with ROC curve.

Table 5
Results of AUC for three algorithms.

Algorithms	AUC	Std. Error	Asymptotic Sig0.	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
DT	0.951	0.025	0.000	0.862	0.99
SVR	0.939	0.032	0.000	0.845	0.984
LR	0.914	0.036	0.000	0.82	0.97

DT is a classification or regression tree depending on whether the target variables are discrete or continuous (Pradhan. 2013). Each tree is made up of nodes and branches that have been broken into multiple child nodes based on the variables that cause the most heterogeneity. Choosing variables that best split the root node into distinct groups is the criterion utilized in the selection procedure (Ramezankhani et al., 2017).

2.4.5. Validation

- Root mean square error (RMSE)

The RMSE estimates the degree of error between two datasets by comparing predicted and measured values. This value is calculated through Eq. 6 (Razavi-Termeh et al., 2021).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \widehat{x}_i)^2}{N}} \tag{6}$$

Where x_i is the values estimated by the model and $\widehat{x}_i$ are the values of the measured data.

- Receiver operating characteristic (ROC) curve

Two axes of sensitivity and specificity make up the ROC curve (Ranjgar et al., 2021). The values of sensitivity and the specificity of the ROC curve were calculated based on Eq. 7 and Eq. 8.

$$X = 1 - \left[\frac{TN}{TN + FP}\right] \tag{7}$$

$$Y = \left[\frac{TP}{TP + FN}\right] \tag{8}$$

The AUC, or area under the ROC curve, represents the degree of system prediction by characterizing its capability to accurately estimate whether something happened or not (Shogrkhodaei et al., 2021). TN stands for true negative, TP for true positive, FN for false negative, and FP for false positive in these equations.

3. Results and discussion

3.1. Multicollinearity of environmental factors

In multicollinearity analysis, the variance coefficient (VIF) should be less than 5 (Razavi-Termeh et al., 2020). According to the VIF, the values of all parameters were less than 5, so there is no collinearity between the independent variables, and all of these variables can be used in modelling (Table 1).

3.2. Spatial autocorrelation of cutaneous leishmaniasis

Moran's I and Getis-ord Gi* indices were utilized to analyze spatial autocorrelation and leishmaniasis hot spot analysis in this research (Table 2). In 2016 and 2017, the assumption that disease distribution is random was accepted based on the Moran index (P-value > 0.05). According to the results, the distribution of disease in the region in 2011 to 2015 and 2018 is cluster. The results of p-value statistics show the significance of the results of spatial autocorrelation.

The results of Getis-ord Gi * index showed that the hot spots of the disease are studied in the northern and central of the region, and the cold spots are located in the west and southwest of the region (Fig. 5). In the northern and central boundaries of Isfahan province, the counties of Aran and Bidgol, Ardestan and Kashan have a cluster distribution of leishmaniasis, where mild and slightly dry winds and temperatures between 28 and 40° C are important factors in the diseases prevalence.

There is no leishmaniasis in the eastern and southern regions of Isfahan province. Khor, Biabank, and Nain are in the province's eastern part, with an average annual temperature of 50 °C, while Semirom is in the province's southernmost part, as one of Iran's highest towns, with an altitude of more than 2460 m and an annual temperature of 10.6 °C. Therefore, considering that temperatures below 30 °C, as well as low-lying lands, are favorable for the outbreak of leishmaniasis (Tsirigotakis et al., 2018), it can be concluded that excessive temperature rise in eastern regions and altitude increase in the southern regions of the province has had an adverse effect on the prevalence of leishmaniasis.

3.3. Fuzzification of environmental factors

Fig. 6 shows the fuzzy functions used in the fuzzification of effective environmental factors. The maps affecting the prevalence of leishmaniasis were used as the input of fuzzy functions. Higher fuzzy values indicate the desirability of more areas in the incidence of leishmaniasis, and the closer they get to zero, the less likely they are to occur.

The prevalence of leishmaniasis rises in warmer climates because the epidemic's vectors prefer to live in tropical and subtropical locations with temperatures between 40 and 50 °C (El Omari et al., 2020). Accordingly, in this study, the temperature of 27 to 30 °C was considered as the optimal temperature because the maximum probability of leishmaniasis incidence is in this temperature range. For this criterion, the Gaussian function was used. Thus, temperatures below 27 °C had the least effect, and temperatures above 30 °C reduced the prevalence of the disease. Since the prevalence of leishmaniasis is directly related to temperature and sunlight causes an increase in temperature, so in this study, the effect of the sun on the prevalence of leishmaniasis was considered positive. The Large fuzzy function was selected, and when the number of sunny days is higher than 255 days, it shows the increasing trend of the leishmaniasis. Enhancement in the number of frosty days led to a decline in temperature and was considered a leishmaniasis spread reducing factor in this study. Therefore, the Small fuzzy function was used and scopes of more than 45 days were considered as the least dangerous areas for leishmaniasis. Raise in the rate of rainfall and humidity reduces the activity of sandflies and consequently reduces the prevalence of leishmaniasis. Thus, in this study, Small functions were used to evaluate the effectiveness of each of these parameters, and areas with less than 44 mm rainfall and less than 38% humidity have the greatest impact on the leishmaniasis risk. Because the optimal altitude for leishmaniasis sandflies is 200 m above sea level, there is little chance of seeing them at high altitudes (Tsirigotakis et al., 2018). An increase in the slope of the region has an indirect relationship with the prevalence of leishmaniasis (Ali Hanafi-Bojd et al., 2015). Therefore, in this study, the Small function was selected. Areas with an altitude more than 591 m above sea level and a slope above 30° were chosen as areas with the lowest number of vectors of leishmaniasis. Vegetation is a parameter that reduces leishmaniasis vectors, and increasing it reduces leishmaniasis prevalence (Bavia et al., 2005). Using the Small fuzzy function, scopes with less than 0.5 vegetation are considered high-risk locations for the presence of vectors in this study. The incidence of leishmaniasis decreases as the wind speed increases. Ramezankhani et al. (2017) found that at speeds greater than 14 m/s, the risk of leishmaniasis decreases. Thus, in the study area, areas with wind speed less than 14 m/s were considered suitable places for the growth and proliferation of reservoirs and disease vectors. The Small fuzzy function was used for the wind speed. There are more disease vector sandflies at distances less than 1 km from the streambed (Tsirigotakis et al., 2018), thus in this study, according to the Small fuzzy function, distances less than 1 km have the greatest impact on the risk of leishmaniasis. Fig. 7 shows the fuzzy maps of environmental parameters.

3.4. Modelling

Fuzzy maps of environmental criteria and leishmaniasis data were considered as modeling input values. 70% of the total data was utilized for training, with the remaining 30% being used for validation. The RMSE index results are summarized in Table 3. The training data fitting error and validation vs. target data diagram are shown in Fig. 8.

The RMSE index for DT revealed a 0.228 error for training data and a 0.289 error for validation data. An example of a DT is illustrated in Fig. 9. Based on the gain ratio in the DT algorithm, the most important factors are rainfall, NDVI, frosty days, altitude, wind speed, sunny day, temperature, slope, humidity, and distance from the stream, respectively. The root in the DT algorithm is the criterion with the largest gain ratio (rainfall), and the gain ratio is computed in each step to determine the leaf. The RMSE value for the SVR algorithm designed to model leishmaniasis at the training stage was 0.274, and the value was 0.324 for validation. The RMSE value for LR algorithm designed in the training phase is 0.292, and the value in the validation phase is 0.347.

In comparison to the other two algorithms, the DT algorithm has a lower RMSE value for both test and training data. The results also showed that the SVR algorithm is more accurate than the LR algorithm in estimating and modeling leishmaniasis in the study area. The SVR algorithm built with kernel function is very flexible and it could detect nonlinear and complex relationships between input and output data (Awad and Khanna, 2015). Table 4 illustrates the unstandardized and standardized coefficients on the effect of independent variables on the dependent variable in the LR algorithm. Wind speed, frosty days, and temperature are all statistically significant and should be factored into the model. This means that each of the independent variables can predict the disease variable. It is also clear from the results of the table that the altitude will have a more effective role in predicting the dependent variable of leishmaniasis.

3.5. Preparation cutaneous leishmaniasis risk mapping (CLRM)

The modeling findings of machine learning algorithms were exported to ArcGIS 10.3 and then applied to the entire study area to create the spatial prediction map of leishmaniasis (Fig. 10). Natural Breaks' categorization technique was used to divide prediction maps into five risk classifications, ranging from very low to very high.

According to the findings, the northern and central regions of Isfahan province have the highest risk of leishmaniasis in all three machine learning algorithms, while the eastern and southern parts of the province have the lowest risk. The climate of the southern regions of Isfahan province is temperate and with cold winters so that sometimes the temperature of the regions is 3.5 °C. According to the results, the southern regions of Isfahan province are the least dangerous regions in terms of the prevalence of leishmaniasis in all three algorithms used. It can be concluded that the temperature decrease in these areas had a high impact on reducing the prevalence of leishmaniasis. The central regions of Isfahan province with semi-desert and dry climates are in the area with an increased risk of leishmaniasis. In these areas, rainfall occurs rarely, humidity is very low, and the temperature is higher than the temperature in the western regions of the province with cold mountain climates. Rising temperatures in these central areas have increased the risk of leishmaniasis outbreaks.

3.6. Validation of cutaneous leishmaniasis risk maps

Evaluating the leishmaniasis prediction maps with three machine learning algorithms by ROC curve is shown in Fig. 11 and Table 5. For the DT, SVR, and LR algorithms, the AUC value indicated the accuracy of 0.951, 0.934, and 0.914, respectively. The DT algorithm had more accuracy in modeling leishmaniasis than the other two algorithms, according to the results of the RMSE and ROC indices. The advantages of this method include simplifying complex relationships between inputs, easy interpretation and control, and management of missing value data, which are commonly encountered in medical studies. In contrast, it is essential to prepare most of the data and control the missing values properly in regression algorithms (Tanner et al., 2008).

LR algorithms are only able to detect linear relationships between independent and dependent variables, and if there are nonlinear relationships between these variables, the efficiency of these algorithms is significantly reduced. As a result, it appears that the nonlinear relationship between the feature analyzed in this study and the inputs utilized in modeling is the main reason for the LR algorithm's poor performance (Lee and Brooks, 2006). Also, SVR made with kernel function has high flexibility and a good ability to detect nonlinear and complex relationships between input and output data. Therefore, according to the output results of the algorithms, it seems that in problems where there are high environmental complexities, and there are complex nonlinear relationships between input and output data and at the same time there is a few available data, the SVR is superior to LR algorithm (Shirzadi et al., 2017).

4. Conclusion

Since Isfahan province is located in the center of the Iranian plateau in fields parallel to the Zayandehrud, which includes mountainous and plain areas, the regions of this province have different altitudes above sea level, which has created a variety of climates, and given that leishmaniasis is a function of various environmental factors, its prediction is very complex. Identifying areas with a high risk of leishmaniasis is the most important issue for most decision-makers and managers.

Based on the fuzzy results, low NDVI, low rainfall, low humidity, low wind speed, high temperature, reduced altitude, and slope had a greater impact on the prevalence of leishmaniasis in the study area. Assessing the accuracy of maps generated using three machine learning algorithms (SVR, DT, LR) showed that DT is the most accurate algorithm for predicting leishmaniasis. The findings revealed that leishmaniasis is a high-risk area in Isfahan province's northern and central areas, but a low-risk area in the province's western areas. Therefore, by creating leishmaniasis prediction maps, the process of accurate and timely information can be accelerated and improved. It is recommended that socioeconomic factors be included in conjunction with environmental factors in future research of cutaneous leishmaniasis prevalence risk modeling.

Funding

This research was supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program (Project No. P0016038).

CRediT authorship contribution statement

Negar Shabanpour: Conceptualization, Data curation, Methodology, Software, Validation, Visualization, Writing – original draft. **Seyed Vahid Razavi-Termeh:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing – original draft. **Abolghasem Sadeghi-Niaraki:** Conceptualization, Methodology, Resources, Software, Supervision, Validation, Writing – review & editing. **Soo-Mi Choi:** Funding acquisition, Project administration, Writing – review & editing. **Tamer Abuhmed:** Resources, Supervision, Validation, Visualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adegboye, O., Adegboye, M., 2017. Spatially Correlated Time Series and Ecological Niche Analysis of Cutaneous Leishmaniasis in Afghanistan. Int. J. Environ. Res. Public Health 14 (3), 309. https://doi.org/10.3390/ijerph14030309.

Akhavan, P., Karimi, M., Pahlavani, P., 2014. Risk Mapping of Cutaneous Leishmaniasis via a Fuzzy C Means-based Neuro-Fuzzy Inference System. Int. Arch. Photogrammetry, Remote Sens. Spatial Inform. Sci..

Alexopoulos, E.C., 2010. Introduction to multivariate regression analysis. Hippokratia. 14 (Suppl 1), 23–28.

Anselin, L., 1995. Local indicators of spatial association—LISA. Geogr. Analy. 27 (2), 93–115.

Awad, M., Khanna, R., 2015. Support Vector Regression. Efficient Learning Machines. Apress, Berkeley, CA.

Battineni, G., Sagaro, G.G., Chinatalapudi, N., Amenta, F., 2020. Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis. J. Pers. Med. 10 (2), 21. https://doi.org/10.3390/jpm10020021.

Bavia, M.E., Carneiro, D.D., Gurgel Hda, C., Madureira Filho, C., Barbosa, M.G., 2005. Remote Sensing and Geographic Information Systems and risk of American visceral leishmaniasis in Bahia, Brazil. Parassitologia. 47 (1), 165–169.

de Santana Martins Rodgers, M., Bavia, M.E., Fonseca, E.O.L., Cova, B.O., Silva, M.M.N., Carneiro, D.D.M.T., Cardim, L.L., Malone, J.B., 2019. Ecological niche models for sand fly species and predicted distribution of Lutzomyia longipalpis (Diptera: Psychodidae) and visceral leishmaniasis in Bahia state, Brazil.

El Omari, H., Chahlaoui, A., Talbi, F.Z., Mouhdi, K.E., El Ouali Lalami, A., 2020. Impact of Climatic Factors on the Seasonal Fluctuation of Leishmaniasis Vectors in. Can. J. Infect. Dis. Med. Microbiol. 2020, 1–7.

Gherbi, R., Bounechada, M., Latrofa, M.S., Annoscia, G., Tarallo, V.D., Dantas-Torres, F., Otranto, D., 2020. Phlebotomine sand flies and Leishmania species in a focus of cutaneous leishmaniasis in Algeria.

Ali Hanafi-Bojd, A., Yaghoobi-Ershadi, M.R., Haghdoost, A.A., Akhavan, A.A., Rassi, Y., Karimi, A., Charrahy, Z., 2015. Modeling the Distribution of Cutaneous Leishmaniasis Vectors (Psychodidae: Phlebotominae) in Iran: A Potential Transmission in Disease Prone Areas. J. Med. Entomol. 52 (4), 557–565.

Hartley, M.-A., Drexler, S., Ronet, C., Beverley, S.M., Fasel, N., 2014. The immunological, environmental, and phylogenetic perpetrators of metastatic leishmaniasis. Trends Parasitol. 30 (8), 412–422.

Imeni, M., 2020. Fuzzy logic in accounting and auditing. J. Fuzzy Extension Appl. 1 (1), 69–75.

Iliopoulou, P., Tsatsaris, A., Katsios, I., Panagiotopoulou, A., Romaliades, S., Papadopoulos, B., Tselentis, Y., 2018. Risk Mapping of Visceral Leishmaniasis: A Spatial Regression Model for Attica Region. Greece 3 (3), 83. https://doi.org/10.3390/tropicalmed3030083.

Jiang, H., Eastman, J.R., 2000. Application of fuzzy measures in multi-criteria evaluation in GIS. Int. J. Geogr. Inform. Sci. 14 (2), 173–184.

Khamesipour, A., Molaei, S., Babaei-Pouya, N., Moradi-Asl, E., 2020. Cutaneous Leishmaniasis Situation and Predicting the Distribution of Phlebotomus papatasi and P. sergenti as Vectors of Leishmaniasis in Ardabil Province, Iran. Korean J. Parasitol. 58 (3), 229–236.

Khashoggi, B.F., Murad, A., 2020. Issues of healthcare planning and GIS: a review. ISPRS Int. J. Geo-Inf. 9 (6), 352.

Khorram, A., Yousefi, M., Alavi, S.A., Farsi, J., 2015. Convenient Landfill Site Selection by Using Fuzzy Logic and Geographic Information Systems: A Case Study in Bardaskan, East of Iran. Health Scope. 4 (1), e19383.

Kohli, P.S., Arora, S., 2018. Application of Machine Learning in Disease Prediction. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1–4.

Länger, B.M., Pou-Barreto, C., González-Alcón, C., Valladares, B., Wimmer, B., Torres, N. V., 2012. Modeling of leishmaniasis infection.dynamics: novel application to the design of effective therapies. BMC Syst. Biol. 6 (1) https://doi.org/10.1186/1752-0509-6-1.

Lee, B.C., Brooks, D.M., 2006. Accurate and efficient regression modeling for microarchitectural performance and power prediction. ACM SIGOPS Operat. Syst. Rev. 40 (5), 185–194.

Marcoulides, K.M., Raykov, T., 2019. Evaluation of Variance Inflation Factors in Regression Models Using Latent Variable Modeling Methods. Educ. Psychol. Measur. 79 (5), 874–882.

Mathur, M., 2015. Spatial autocorrelation analysis in plant population: An overview. J. Appl. Nat. Sci. 7 (1), 501–513.

Melo, H.A., Rossoni, D.F., Teodoro, U., 2018. Effect of vegetation on cutaneous leishmaniasis in Paraná, Brazil.

Mohammadbeigi, A., Khazaei, S., Heidari, H., Asgarian, A., Arsangjang, S., Saghafipour, A., Mohammadsalehi, N., Ansari, H., 2020 Sep 7. An investigation of the effects of environmental and ecologic factors on cutaneous leishmaniasis in the old world: a systematic review study. Rev. Environ. Health 36 (1), 117–128.

Mohebali, M., Keshavarz, H., Shirmohammad, S., Akhoundi, B., Borjian, A., Hassanpour, G., Mamishi, S., Mahmoudi, S., 2020. The diagnostic accuracy of direct agglutination test for serodiagnosis of human visceral leishmaniasis: a systematic review with meta-analysis. BMC Infect. Dis. 20 (1) https://doi.org/10.1186/s12879-020-05558-7.

Mollalo, A., Sadeghian, A., Israel, G.D., Rashidi, P., Sofizadeh, A., Glass, G.E., 2018. Machine learning approaches in GIS-based ecological modeling of the sand fly Phlebotomus papatasi, a vector of zoonotic cutaneous leishmaniasis in Golestan province, Iran. Acta Trop. 188, 187–194.

Moradiasl, E., Rassi, Y., Hanafi-Bojd, A., Vatandoost, H., Saghafipour, A., Adham, D., Aabasgolizadeh, N., Omidi Oskouei, A., Sadeghi, H., 2018. The Relationship between Climatic Factors and the Prevalence of Visceral Leishmaniasis in North West of Iran. Int. J. Pediatrics 6 (2), 7169–7178.

Nilforoushzadeh, M.A., Shirani Bidabadi, L., Hosseini, S.M., Fadaei Nobari, R., Jaffary, F., Cutaneous Leishmaniasis in Isfahan Province, Iran, During 2001-2011, J. Skin Stem. Cell.

Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput. Geosci. 51, 350–365.

Rajabi, M., Pilesjö, P., Shirzadi, M.R., Fadaei, R., Mansourian, A., 2016. A spatially explicit agent-based modeling approach for the spread of Cutaneous Leishmaniasis disease in central Iran, Isfahan. Environ. Modell. Software 82, 330–346.

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

Ranjgar, B., Razavi-Termeh, S.V., Foroughnia, F., Sadeghi-Niaraki, A., Perissin, D., 2021. Land subsidence susceptibility mapping using persistent scatterer SAR interferometry technique and optimized hybrid machine learning algorithms. Remote Sensing 13 (7), 1326. https://doi.org/10.3390/rs13071326.

Razavi-Termeh, S.V., Sadeghi-Niaraki, A., Choi, S.-M., 2021. Effects of air pollution in Spatio-temporal modeling of asthma-prone areas using a machine learning model. Environ. Res. 200, 111344. https://doi.org/10.1016/j.envres.2021.111344.

Razavi-Termeh, S.V., Khosravi, K., Sadeghi-Niaraki, A., Choi, S.-M., Singh, V.P., 2020. Improving groundwater potential mapping using metaheuristic approaches. Hydrol. Sci. J. 65 (16), 2729–2749.

Razavi-Termeh, S.V., Sadeghi-Niaraki, A., Choi, S.M., 2022. Spatio-temporal modeling of asthma-prone areas using a machine learning optimized with metaheuristic algorithms. *Geocarto Int.* (just-accepted), 1-22.

Sharma, P., Singh Chandel, V.B., Kahlon, S., 2018. Spatial Autocorrelation Technique for Landslide Hot- spot Analysis in the Upper Ravi River Catchiment, Chamba, Himachal Pradesh. Indian Geogr. J. 93(1), pp. 1-9, ISSN 0019-4824.

Shirzadi, A., Shahabi, H., Chapi, K., Bui, D.T., Pham, B.T., Shahedi, K., Ahmad, B.B., 2017. A comparative study between popular statistical and machine learning methods for simulating volume of landslides. CATENA 157, 213–226.

Shogrkhodaei, S.Z., Razavi-Termeh, S.V., Fathnia, A., 2021. Seyed Vahid Razavi-Termeh, and Amanollah Fathnia. "Spatio-temporal modeling of pm2. 5 risk mapping using three machine learning algorithms.". Environ. Pollut. 289, 117859. https://doi.org/10.1016/j.envpol.2021.117859.

Sofizadeh, A., Rassi, Y., Vatandoost, H., Hanafi-Bojd, A.A., Mollalo, A., Rafizadeh, S., Akhavan, A.A., 2017 Mar 1. Predicting the Distribution of Phlebotomus papatasi (Diptera: Psychodidae), the Primary Vector of Zoonotic Cutaneous Leishmaniasis, in Golestan Province of Iran Using Ecological Niche Modeling: Comparison of MaxEnt and GARP Models. J. Med. Entomol. 54 (2), 312–320.

Tabasi, M., Alesheikh, A.A., Sofizadeh, A., Saeidian, B., Pradhan, B., AlAmri, A., 2020. A spatio-temporal agent-based approach for modeling the spread of zoonotic cutaneous leishmaniasis in northeast Iran. Parasites Vectors 13 (1). https://doi.org/10.1186/s13071-020-04447-x.

Tanner, L., Schreiber, M., Low, J.G.H., Ong, A., Tolfvenstam, T., Lai, Y.L., Ng, L.C., Leo, Y.S., Thi Puong, L.e., Vasudevan, S.G., Simmons, C.P., Hibberd, M.L., Ooi, E.E., Farrar, J., 2008. Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. PLoS Negl. Trop. Dis. 2 (3), e196.

Tsirigotakis, N., Pavlou, C., Christodoulou, V., Dokianakis, E., Kourouniotis, C., Alten, B., Antoniou, M., 2018. Phlebotomine sand flies (Diptera: Psychodidae) in the Greek Aegean Islands: ecological approaches. Parasites Vectors 11 (1), 1–14.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Netw. 10 (5), 988–999.

VoPham, T., Hart, J.E., Laden, F., Chiang, Y.Y., 2018. Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. Environ. Health 17 (1), 1–6.

Waitz, Y., Paz, S., Meir, D., et al., 2019. Effects of land use type, spatial patterns and host presence on Leishmania tropica vectors activity. Parasites Vectors 12, 320.

WHO, 2012. A human rights-based approach to neglected tropical diseases. World Health Organization, Geneva. http://www.who.int/tdr/publications/tdrresearchpublications/human-ights/en/index.html (accessed 3 September 2012).

Wu, C.-C., Yeh, W.-C., Hsu, W.-D., Islam, M.M., Nguyen, P.A., Poly, T.N., Wang, Y.-C., Yang, H.-C., (Jack) Li, Y.-C., 2019. Prediction of fatty liver disease using machine learning algorithms. Comput. Methods Programs Biomed. 170, 23–29.

Yariyan, P., Avand, M., Soltani, F., Ghorbanzadeh, O., Blaschke, T., 2020. Earthquake vulnerability mapping using different hybrid models. Symmetry 12 (3), 405.

Yashon O. Ouma, Clinton O. Okuku, Evalyne N. Njau, 2020. Use of Artificial Neural Networks and Multiple Linear Regression Model for the Prediction of Dissolved Oxygen in Rivers: Case Study of Hydrographic Basin of River Nyando, Kenya, Complexity, vol. 2020, Article ID 9570789, 23 p.

Further reading

de Almeida-Pititto, B., Dualib, P.M., Zajdenverg, L., Dantas, J.R., de Souza, F.D., Rodacki, M., Bertoluci, M.C., 2020. Severity and mortality of COVID 19 in patients with diabetes, hypertension and cardiovascular disease: a meta-analysis. Diabetol. Metab. Syndr. 12 (1) https://doi.org/10.1186/s13098-020-00586-4.