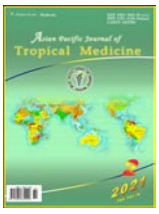

Original Article

Asian Pacific Journal of Tropical Medicine

apjtm.org

doi: 10.4103/1995-7645.306739

Impact Factor: 1.94

Predicting cutaneous leishmaniasis using SARIMA and Markov switching models in Isfahan, Iran: A time-series study

Vahid Rahmanian[1], Saied Bokaie[1]✉, Aliakbar Haghdoost[2], Mohsen Barouni[3]

[1]*Department of Food Hygiene and Quality Control, Division of Epidemiology & Zoonoses, Faculty of Veterinary Medicine, University of Tehran, Tehran, Iran*

[2]*HIV/STI Surveillance Research Center, and WHO Collaborating Center for HIV Surveillance, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran*

[3]*Health Services Management Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran*

ABSTRACT

Objective: To determine the potential effect of environment variables on cutaneous leishmaniasis occurrence using time-series models and compare the predictive ability of seasonal autoregressive integrated moving average (SARIMA) models and Markov switching model (MSM).

Methods: This descriptive study employed yearly and monthly data of 49 364 parasitologically-confirmed cases of cutaneous leishmaniasis in Isfahan province, located in the center of Iran from January 2000 to December 2019. The data were provided by the leishmaniasis national surveillance system, the meteorological organization of Isfahan province, and Iranian Space Agency for vegetation information. The SARIMA and MSM models were implemented to examine the environmental factors of cutaneous leishmaniasis epidemics.

Results: The minimum relative humidity, maximum relative humidity, minimum wind speed, and maximum wind speed were significantly associated with cutaneous leishmaniasis epidemics in different lags (P<0.05). Comparing SARIMA and MSM, Akaikes information criterion (AIC), and mean absolute percentage error (MAPE) in MSM were much smaller than SARIMA models (MSM: AIC=0.95, MAPE=3.5%; SARIMA: AIC=158.93, MAPE:11.45%).

Conclusions: SARIMA and MSM can be a useful tool for predicting cutaneous leishmaniasis in Isfahan province. Since cutaneous leishmaniasis falls into one of two states of epidemic and non-epidemic, the use of MSM (dynamic) is recommended, which can provide more information compared to models that use a single distribution for all observations (Box-Jenkins SARIMA model).

KEYWORDS: Leishmaniasis; Climate factor; Time series analysis; Forecasting; Iran

1. Introduction

Leishmaniasis is a neglected vector born disease caused by the protozoan parasites of genus *Leishmania* that occurs in three forms, *i.e.* cutaneous, mucosal, and visceral forms[1]. Cutaneous leishmaniasis (CL) has been reported with an approximately annual incidence between 0.7-1.2 per million new cases worldwide. More than 90% of the CL cases have occurred in 8 countries, *i.e.* Afghanistan, Algeria, Brazil, Peru, Saudi Arabia, Syria, Iraq, and Iran[2].

According to the Ministry of Health and Medical Education of Iran, annually, about 30 000 cases of CL occurred in Iran. However, the number of actual cases is estimated to be four to five times[3]. There are two types of CL namely anthroponotic (ACL) and zoonotic CL (ZCL) in Iran. The ACL type is caused by *Leishmania (L.) tropica*, whose main reservoir is human and the vector is *Phlebotomus (P.) sergenty*[4]. The ZCL type is caused by *L. major* and the rodents including *Rhombomys opimus*, *Meriones libycus*, and *Meriones nesokia* are considered as the main reservoir. The main vector of the ZCL is *P. papatasi*[5,6]. This type is the most prevalent and has been found in 25 of 31 provinces in Iran, including Isfahan province as the oldest endemic foci of ZCL with the highest incidence in Iran[7].

Notwithstanding, the extensive public health activities have focused on prevention and control in surveillance systems; hence

✉To whom correspondence may be addressed. E-mail: sbokaie@ut.ac.ir

How to cite this article: Rahmanian V, Bokaie S, Haghdoost A, Barouni M. Predicting cutaneous leishmaniasis using SARIMA and Markov switching models in Isfahan, Iran: A time-series study. Asian Pac J Trop Med 2021; 14(2): 83-93.

Article history: Received 2 September 2020 Revision 26 November 2020
Accepted 10 December 2020 Available online 20 January 2021

it was found that Isfahan had the highest mean incidence of CL between 1983 and 2013[3,4]. Several studies have reported the relation between CL and environmental factors such as rainfall, air temperature, slope of area, wind speed, evaporation rate, relative humidity, and vegetation cover[8-10]. A few studies have considered the effect of climate and environmental factors in time series analysis for modelling and prediction of CL in Iran.

Establishing a monitoring and surveillance system is necessary for both rapid detections of leishmaniasis outbreaks and for the monitoring of possible epidemics. Also, the public health system, especially in endemic areas, should be fully prepared to plan for potential outbreaks. Therefore, it is very important to use the models to predict the occurrence of leishmaniasis, especially in areas with high incidence. One of the most common models used in epidemiology is data analysis of time series. The purpose of using this method is to predict future values and the factors influencing its occurrence over time. Different models in time series analysis such as Gray models, general regression model, Box-Jenkins models, and negative binomial regression and neural network have been used to describe the pattern and predict the incidence of infectious diseases[11-13]. Each of these has its own considerations and accuracy.

In recent years, Box-Jenkins seasonal autoregressive integrated moving average (SARIMA) models have been used in medicine. Because infectious diseases fall into one of two stages of epidemic or non-epidemic and the use of these methods for early detection of outbreaks has limitations[14]. The use of mixed models in this case seems to be superior to models that use a single distribution for all observations[15]. Therefore, the objectives of this study are: (1) to determine the potential effect of environment variables on CL occurrence using time-series models, and (2) to compare the predictive ability of Markov switching model (MSM) and Box-Jenkins SARIMA models.

2. Materials and methods

2.1. Study site

Isfahan province is located in the center of Iran in the Zayandeh-Rud River of green plain, at the foothills of the Zagros mountain. The province had a population of 5 120 850 (4 507 309 urban and 613 073 rural area) according to General Population and Housing Census in 2016. The ancient and major city of Isfahan is the center of the province, making it the third-largest city in Iran. It also consists of 25 counties with an average altitude of 1 590 meters above sea level. The annual average of the minimum and maximum temperatures in this province is 10 ℃ and 25 ℃, respectively. The climate in this area is highly variable so that the average annual rainfall is between 100 and 150 mm. The air humidity is in the range of 11% to 82% and the average number of hours of sunshine per month is 260 hours.

2.2. Data collection

This descriptive study employed yearly and monthly data of 49 364 parasitologically-confirmed cases of CL in Isfahan province between January 2000 and December 2019. The data were provided by the leishmaniasis national surveillance system at Isfahan University of Medical Sciences.

All climate and weather data such as the average monthly temperature (℃), average monthly maximum and minimum temperature (℃), average monthly pressure (p), average monthly relative humidity (%), average maximum and minimum relative humidity (%), total hours of sunshine per month, total number of rainy days per month, total monthly rainfall (mm), number of days with dust, maximum and minimum monthly wind speed (km/h) were gathered from the meteorological organization of Isfahan province. The data were collected from 10 synoptic centres in Isfahan province, which shows the highest incidence of CL.

Vegetation information was extracted from the Moderate Resolution Imaging Spectroradiometer (16-day composites) satellite, which is accessible to the public. Moderate Resolution Imaging Spectroradiometer uses remote sensing data of the Normalized Difference Vegetation Index (NDVI) for analysis, measurement, and evaluating the presence or absence of vegetation in an area. The range of NDVI changes is between +1 and -1, where negative values of the index (numbers close to -1) indicate water zones. The values close to zero (between -0.1 and +0.1) usually indicate bare rock, sand, or snow surfaces. Low and positive values of the index (about +0.2 to +0.4) indicate vegetation of shrubs and grasslands, and high values of NDVI (numbers close to +1) indicate dense vegetation (dense green forests)[16]. The NDVI data of the province were accessible *via* the website of the Iranian Space Agency for the period between January 2000 and December 2019[17]. Before choosing any model, the trend of CL incidence and environmental variables were observed by plotting data and these plots showed the seasonal trend with periodicity. Also, due to the sinusoidal wave of the coefficients and the significance of the correlation in the lags of 12, 24, and 36 months, it seems that the occurrence of CL and environmental variable does have a seasonal pattern. Therefore, the SARIMA model was selected.

2.3. Statistical analysis

2.3.1. SARIMA model

The monthly occurrences case during the period of this study constructed the SARIMA model. This model is described by seven parameters including SARIMA (p,d,q), (P,D,Q), and s where p is the number of autoregressive, d is the number of times the model was differenced, q is the number of moving averages, P is the number of seasonal autoregressive, D is the number of seasonal integration or seasonal differencing, Q is the number of seasonal moving average, and s is the length of seasonal periods[8].

To check the stationary in variance and mean, Box-Cox regression

and Dickey-Fuller test were applied, respectively. To eliminate seasonality trend, the first-order seasonal differencing (D=1) with a period of 12 was used. For estimation of the moving average and autoregressive parameters, autocorrelation functions (ACF) and partial autocorrelation functions (PACF) were identified[18].

In the next step, Akaike information criteria (AIC) lower, Bayesian information criterion (BIC), and likelihood ratio test among possible models showed a better approach among models which were expanded by different lags[15,19]. Finally, a SARIMA (1,0,1) (1,1,1) 12 model was selected with the lower AIC (197.35) and BIC (225.17) values to fit the data in the best way.

Independent variables as external regress with different lags were observed in relation to CL cases using cross-correlation coefficients to detect the best predictor; hence its best lags were included in the final multi-regression SARIMA model.

Later, the achieved SARIMA model was applied to predict the frequency of monthly occurrences case. The forecasting precision was estimated by the mean absolute percentage error (MAPE), which was computed using equation (1). The lower the value, the higher the accuracy of the fitted model.

$$MAPE = \frac{1}{N}\sum_{t=1}^{n}\frac{\text{Actual cases - Predicted cases}}{\text{Actual cases}} \quad (1)$$

Where N is the number of predictions.

Furthermore, to measure the goodness-of-fit of the final model, the portmanteau test was run for white noise and the normality of the residuals was observed on the graphs[19].

2.3.2. MSM

The simple time series models cannot explain nonlinear occurrence of outcomes such as the data of surveillance system of infectious diseases, therefore the MSM model is used due to the problem of structural failure and nonlinearity of series[15]. MSM model is one of the most popular for nonlinear time series models that shows the occurrence of outcomes in different states[14].

Since the CL series has a switching state from epidemic and non-epidemic and *vice versa* (as the name of the switching suggests), therefore this model was found suitable for analysing and predicting such data. In this study, modeling is done separately for the epidemic and non-epidemic state.

A simple MSM model can be written as follows (in this model, S_t=1 means epidemic state and S_t=0 means non-epidemic state):

$$yt = a_{0.0} + a_{0.1}S_t + (a_{1.0} + a_{1.1}S_t)\, y_{t-1} + e_t \quad (2)$$

$$p\,(S_t = j/S_{t-1} = i) = p_{ij} \quad (3)$$

$$S_t \in \{0.1\} \quad (4)$$

$$e_t \sim N\,(0.\sigma^2) \quad (5)$$

This model shows that in the non-epidemic state S_t=0 and the value of y_t is determined based on $a_{0.0}$ constant and the autoregressive parameter with $a_{0.1}$. If an epidemic occurs (S_t=1), the constant value increases to $a_{0.0}+a_{0.1}$ and the autoregressive parameter value increases to $a_{1.0}+a_{1.1}$ (Equation 2).

Equation 3 shows that the hidden states S_t is a function of Markov significant with the transition probability. P_{ij} is the probability of state j at time t conditional to state i at time $t-1$:

$$p''_{ij}\,(s_t = \frac{j}{s_{t-1}} = i.S_{t-2} = k....S_{t-k} = n) = p\,(s_t = \frac{j}{s_{t-1}} = i) = p''_{ij} \quad (6)$$

$$p''_{ij} + p + ... + p_{in} = 1'' \quad (7)$$

The transition matrix is also an N*N matrix consisting of P_{ij}. After fitting the model and estimating the parameters, the values of p_{00} and p_{10} were also estimated. According to the values of these parameters, the matrix of transition probabilities can be formed as follows:

$$p'' = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \begin{bmatrix} p_{00} & 1\text{-}p_{00} \\ 1\text{-}p_{11} & p_{11} \end{bmatrix}'' \quad (8)$$

In this matrix, the sum of the probabilities of each row is equal to one. Here, p_{00} means that the regime is likely to be at non-epidemic state at time t, provided it is at non-epidemic state at time $t-1$. The p_{11} means the probability that the regime will be in a state of epidemic at time t, provided that it is also in a state of epidemic at time $t-1$. The p_{01} means the probability that the regime is in a state of epidemic at time t, provided that it is at non-epidemic at time $t-1$. The means p_{10} the probability that the regime is in a non-epidemic state at time t, provided that it is in a state of epidemic at time $t-1$. To control the seasonality trend, the parametric method of periodic function (Fourier terms) was used. In this method, Sinus and Cosinus functions of time (according to the period of the seasonal pattern based on the data) were entered into the model along with other variables. Also, other independent variables may enter the model with lag or different delays (here environmental variables), in which case the final model is written as follows:

$$Yt = a_{0.0} + a_{0.1}S_t + (a_{1.0} + a_{1.1}S_t)\, y_{t-1} + b_1 \sum_{i=1}^{12}\sin(\tfrac{2\pi i}{12}) + \sum_{i=1}^{12}\cos(\tfrac{2\pi i}{12}) + \sum_{i=1}^{12} civt.i + e_t \quad (9)$$

Where $\sum_{i=1}^{12}\sin(\frac{2\pi i}{12}) + \sum_{i=1}^{12}\cos(\frac{2\pi i}{12})$ are related to the method of seasonality control, which because our data is monthly, the value of i varies from 1 to 12 (January=1, February=2, March=3,…). The expression of $\sum_{i=1}^{12} civt.i$ refers to the entry of independent variables in the model that can enter the model with lag or different delays.

The validity of the MSM was also examined by fitting the data for the period of the study. Nonetheless, the final SARIMA and MSM models were compared based on AIC, BIC, and MAPE indices. Data analysis was performed using Stata software (version 14) and the package time series analysis. The alpha level was 0.05.

2.4. Ethical considerations

The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Faculty of Veterinary Medicine, University of Tehran, Iran (Project identification code: 7265130).

3. Results

3.1. Descriptive analysis

A total of 49 364 cases of CL have been reported in Isfahan province between 2000 and 2019. Figure 1 shows that the trend of CL is seasonal. There is a clear increase in the number of cases from 2011 to 2015, then this starts to decline in 2015 while again it follows an upward trend in 2018. The CL trend is almost stable from January to June. Then it starts to climb from July and reaches its peak in September and October and then starts to decline (Figure 2). Figure 3 shows the trend of significant environmental data between January 2000 and December 2019 in Isfahan province that was used for modelling.

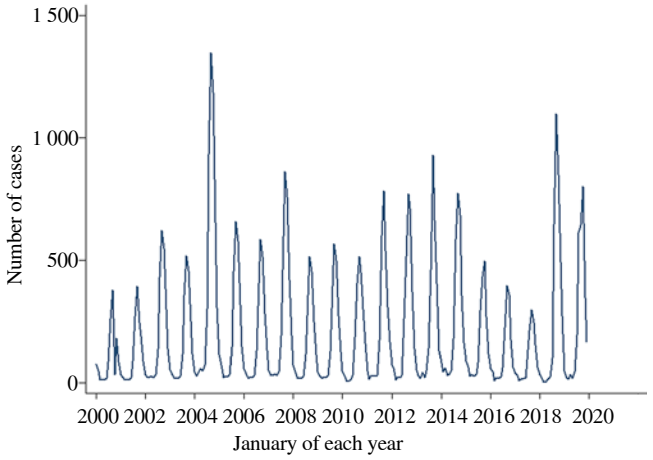

Figure 1. The trend and distribution of 49 364 cutaneous leishmaniasis cases between January 2000 and December 2019 in Isfahan province.

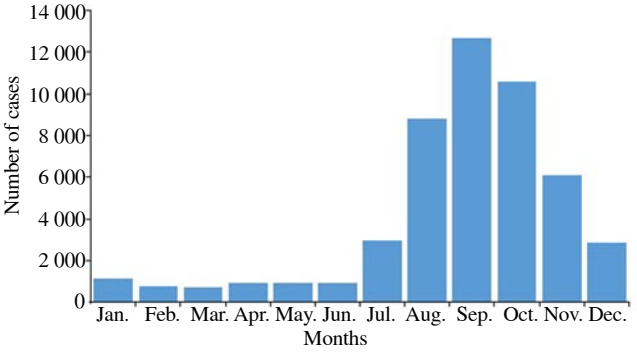

Figure 2. Monthly cumulative distribution of the number of cutaneous leishmaniasis cases in Isfahan province between 2000 and 2019.

3.2. Cross-correlation function

Table 1 shows that there are cross-correlation coefficients and a significant level of CL incident with the average relative humidity at lag of 2 months, maximum relative humidity at lag of 2 and 8 months, minimum relative humidity at lag of 2 months, minimum wind speed at lag of 0 and 8 months, maximum wind speed at lag of 0, 1, 2, and 3 months, total hours of sunshine at lag of 2

Table 1. Cross-correlation coefficients of environmental variables in Isfahan province from zero lag (communication without delay) to 12 months' lag.

Time lag (Months)	Average temperature (℃)	Maximum temperature (℃)	Minimum temperature (℃)	Average relative humidity (%)	Maximum relative humidity (%)	Minimum relative humidity (%)	Average air pressure	Minimum wind speed	Maximum wind speed	Total monthly rainfall	Number of rainy days	Total hours of sunshine	Number of days with dust	Mean NDVI
-12	0.070	0.080	0.050	-0.090	-0.09	-0.020	-0.030	0.010	-0.0100	-0.060	-0.100	0.02	-0.060	0.030
-11	-0.030	-0.005	-0.060	-0.020	-0.05	-0.030	-0.060	0.020	-0.0100	-0.006	-0.030	0.04	-0.090	0.120*
-10	-0.050	-0.050	0.030	-0.010	0.05	-0.030	-0.003	0.090	0.0070	-0.070	-0.007	0.01	0.020	0.030
-9	0.050	0.060	-0.030	-0.080	-0.10	-0.010	-0.004	-0.040	-0.0800	-0.060	-0.090	0.11	-0.020	-0.190*
-8	0.090	0.090	-0.020	-0.070	-0.13*	-0.008	-0.120	0.130*	0.0300	-0.070	-0.040	0.01	-0.001	-0.090
-7	0.080	0.080	0.120	-0.007	-0.07	0.030	-0.100	-0.050	-0.0700	0.020	-0.040	-0.01	0.001*	-0.030
-6	-0.050	-0.060	-0.070	0.090	0.04	0.050	-0.030	-0.050	-0.0004	0.030	-0.002	-0.07	-0.120	0.070
-5	-0.006	-0.020	-0.010	0.060	0.06	0.040	-0.009	0.020	-0.0100	0.070	0.020	-0.10	-0.060	-0.030
-4	-0.010	-0.040	-0.020	0.050	0.07	0.030	0.020	0.002	0.0800	0.100	0.100	-0.04	-0.030	0.050
-3	-0.010	-0.020	-0.003	0.110	0.06	0.060	-0.050	0.010	0.1400*	0.100	0.040	-0.12*	0.110	-0.020
-2	-0.110	-0.110	-0.060	0.160*	0.14*	0.210*	0.070	-0.090	-0.0500*	0.110	0.100	-0.13*	-0.060	-0.060
-1	-0.060	-0.070	-0.110	0.030	0.11	0.001	0.060	0.010	-0.0150*	0.070	0.020	0.04	-0.070	0.001
0	-0.050	-0.060	0.120	0.050	0.02	0.080	0.080	-0.170*	-0.1500*	0.090	0.080	-0.05	-0.050	0.140*

*A level less than 0.05 is significant.

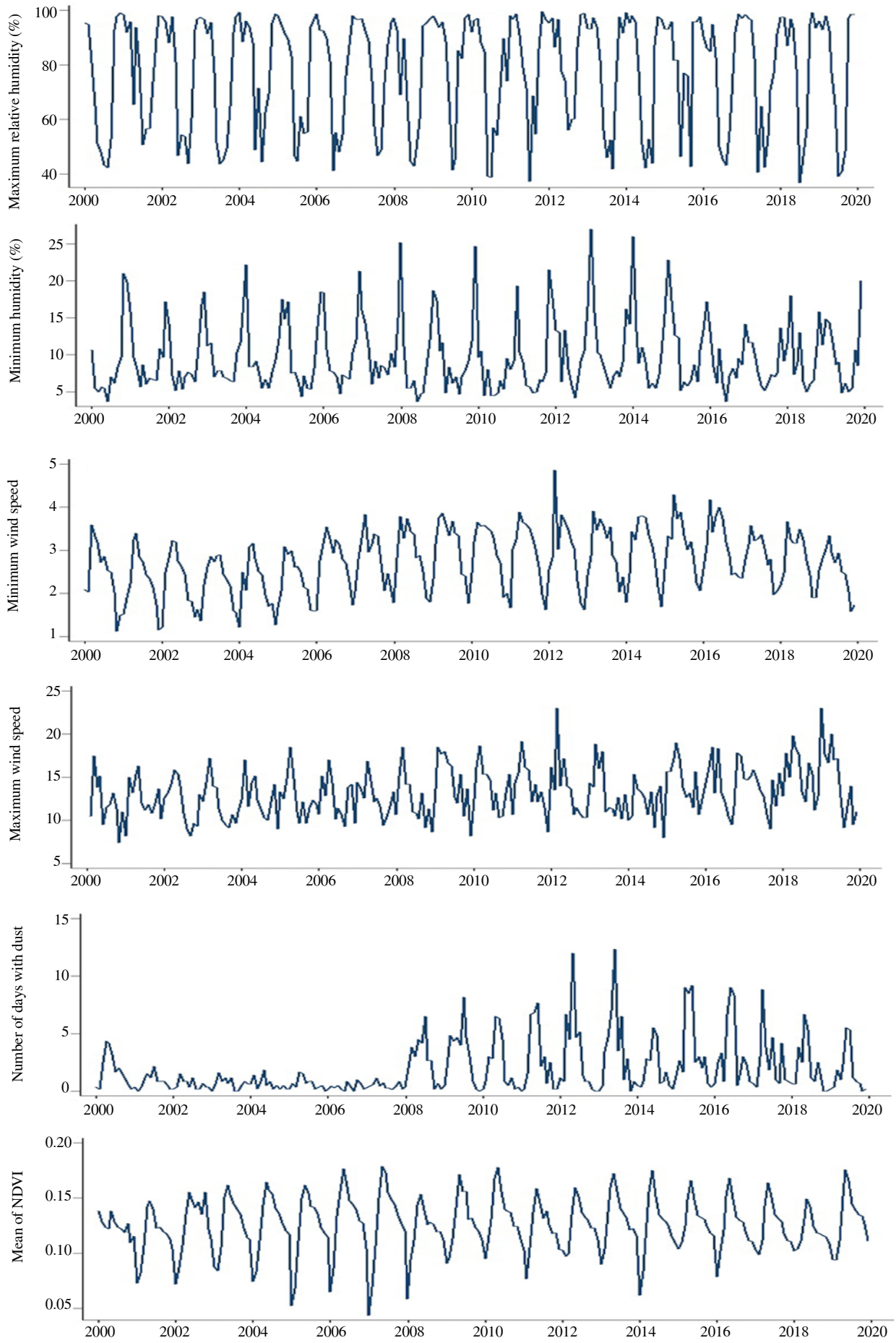

Figure 3. The time series and trend of environmental data between January 2000 and December 2019 in Isfahan province.

and 3 months, number of days with dust at lag of 7 months, mean NDVI at lag of 0, 9 and 11 months. However, there is no significant correlation between other variables and the incidence of CL in any of the lags. Zero lags were not considered due to its insignificance and non-applicability.

In the next step, due to the possibility of collinearity between independent variables in significant lags, Pearson correlation coefficient was used and if there is a strong correlation (above 0.7) between independent variables, one of them was entered the model. There is a strong correlation between the average relative humidity at lag of 2 months with the maximum relative humidity at lag of 2 months (r=0.84), the minimum relative humidity at lag of 2 months (r=0.88), total hours of sunshine at lag of 2 months (r=-0.72) and total hours of sunshine at lag of 3 months (r=-0.92). Therefore, it was decided not to enter the average relative humidity at lag of 2 months in the model. On the other hand, there was a high correlation between the maximum relative humidity at lag of 2 months with total hours of sunshine at lag of 2 months (r=-0.87) and the minimum relative humidity at lag of 2 months with total hours of sunshine at lag of 3 months (r=-0.76). It was decided not to include total hours of sunshine at lag of 2 and 3 months in the model (Table 2). Finally, with the above considerations, the following was entered to the model: maximum relative humidity at lag of 2 and 8 months, minimum relative humidity at lag of 2 months, minimum wind speed at lag of 8 months, maximum wind speed at lag of 1, 2, and 3 months, number of days with dust at lag of 7 months, and the mean NDVI at lag of 9 and 11 months.

Table 2. Pearson correlation coefficient between independent variables in significant lags.

Variable	Pearson correlation coefficient	Significance level
Average relative humidity lag 2		
Maximum relative humidity lag 2	0.84	0.015
Minimum relative humidity lag 2	0.88	0.001
Total hours of sunshine lag 2	-0.72	0.008
Total hours of sunshine lag 3	-0.92	0.005
Maximum relative humidity lag 2 Total hours of sunshine lag 2	-0.87	0.001
Minimum relative humidity lag 2 Total hours of sunshine lag 3	-0.76	0.001
Number of days with dust lag 7		
Minimum wind speed lag 8	0.38	0.010
Mean NDVI lag 9	0.40	0.001
Mean NDVI lag11	0.40	0.001
Maximum relative humidity lag 8 Minimum wind speed lag 8	-0.33	0.001
Minimum relative humidity lag 2		
Maximum wind speed lag 1	-0.21	0.001
Maximum wind speed lag 2	-0.21	0.001
Maximum wind speed lag 3	-0.21	0.001
Maximum relative humidity lag 2	0.61	0.001

The bold fonts indicate reference group.

3.3. SARIMA model

Initially, the above independent variables were entered into the model as univariate, and finally, independent variables with a P-value of less than 0.2 were entered into the model as multivariate. Also, the non-significant variables of multiple models were removed one by one from the model (backward method) and the models were examined with Likelihood ratio test.

Finally, the best-fit model with the lowest AIC and BIC values, the standard error, and 95% confidence interval of coefficients and significance was selected (Table 3).

According to Table 3, CL is significantly associated with the variables of minimum relative humidity at a lag of 2 months, and maximum wind speed at lag of 3 months. The minimum humidity variable with a significantly low value (P=0.002) seems to be the most important variable involved in the fluctuations of CL. The vegetation cover at a lag of 9 months, is present in the model and has the highest coefficient value (-4.64). The presence of this variable in the model improved the AIC but has not shown a significant association with CL incidence (P=0.07).

The AIC and BIC values of model with explanatory variables were less than the model without explanatory variables. Also, the calculated value of MAPE was 11.45%.

To confirm the goodness of fit of the model, the residuals (differences between actual cases and predicted cases) were compared. The histogram of the residues showed that the residues follow the normal distribution. Moreover, the portmanteau test confirmed the residuals normality and independence (P=0.89).

Figure 4 shows the fitted cumulative number of CL cases between January 2000 and December 2019. To evaluate the validity of the model, we showed the fitted model with actual data of CL between January 2000 and December 2019, and then predicted the number of CL cases with 95% confidence intervals between January 2020 and December 2021, based on the final best-fit model. The predicted model was completely fitted to the actual data (Figure 4).

3.4. MSM

In MSM model, similar to the SARIMA model, at the beginning of preparing the series or data to fit the model, Box-Cox regression test was used to check the stationary in variance of the data. This test showed that CL data were not stationary in variance and therefore according to the value of theta (theta=0.08), a log transformation was used. Autocorrelation and partial autocorrelation plots and Dickey-Fuller test ($P \leqslant 0.001$) showed that the data are stationary in the mean.

In this study, because modelling is for an infectious disease, two regimes (state=2) were considered for modelling. State 0 was

Table 3. Parameter coefficients of simple and multiple SARIMA (1,0,1) (1,1,1) 12 model for cutaneous leishmaniasis cases in Isfahan province (seasonal differencing was used for seasonality control).

Parameters	Simple SARIMA model				Multiple SARIMA model			
	Coefficient	S.E	95% *CI*	*P*-value	Coefficient	S.E	95% *CI*	*P*-value
AR	0.88	0.04	0.79, 0.96	<0.001	0.92	0.03	0.85, 0.99	<0.001
MA	-0.40	0.07	-0.054, 0.025	<0.001	-0.37	0.07	-0.50, -0.23	<0.001
SAR	-0.05	0.06	-0.18, 0.06	0.35	-0.13	0.07	-0.29, 0.01	0.07
SMA	-0.88	0.07	-1.02, 0.73	<0.001	-0.78	0.07	-0.93, -0.63	<0.001
Minimum relative humidity-lag 2	-	-	-	-	0.01	0.004	0.005, 0.020	0.002
Minimum wind speed-lag 8	-	-	-	-	0.12	0.07	-0.01, 0.27	0.07
Maximum wind speed-lag 3	-	-	-	-	-0.01	0.008	0.002, 0.030	0.02
Mean NDVI-lag 9	-	-	-	-	-4.64	2.60	-9.75, 0.45	0.07
Constant	0.001	0.02	-0.03, -0.04	0.92	0.02	0.04	-0.06, 0.11	0.60
Sigma	0.33	0.01	0.31, 0.36	<0.001	0.32	0.01	0.029, 0.340	<0.001
AIC	184.43	-	-	-	158.93	-	-	-
BIC	205.00	-	-	-	192.82	-	-	-

AR: Auto-regressive, MA: Moving average, SAR: Seasonal auto-regressive, SMA: Seasonal moving average, "-" not applicable.

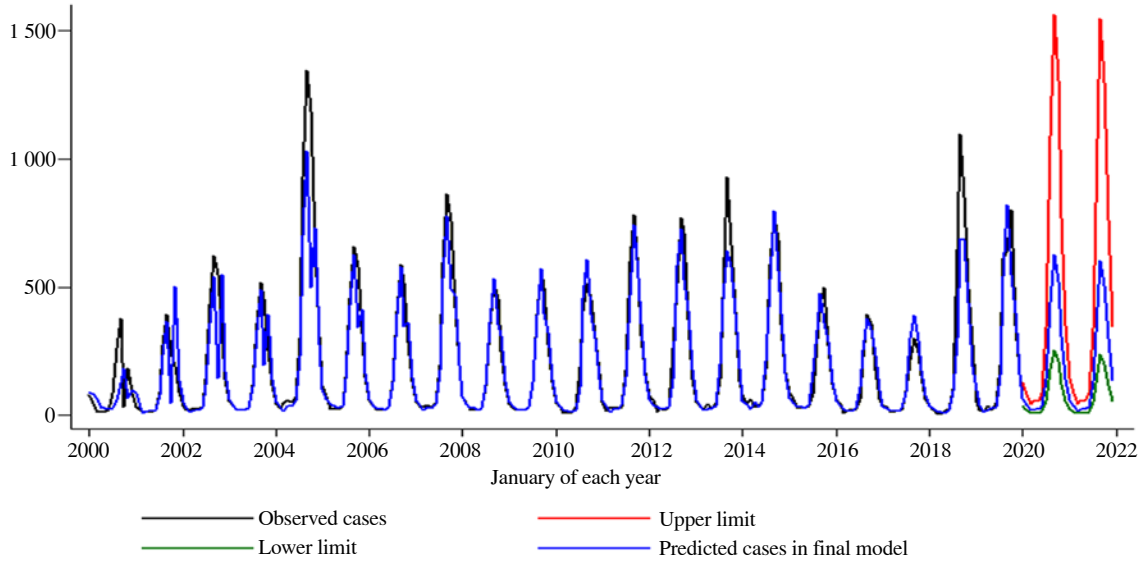

Figure 4. Observed cutaneous leishmaniasis cases between January 2000 and December 2019 and 1-step ahead predicted values between January 2020 and December 2021 based values on the multiple SARIMA (1,0,1) (1,1,1) 12 model with 95% confidence intervals.

considered for the non-epidemic regime and state 1 was for the epidemic regime. Table 4 lists the results of the fitting to MSM model.

If we consider the constant estimated by the model in epidemic state as the epidemic threshold, and according to this model, the number of cases above 108 cases for CL in the province per month can be considered as the outbreak threshold. This is because log transformation was used to stationary the variance of the series, and the exponential of this coefficient was used to reach the real value, so that exp (4.68)=108.12.

Also, as the constant variance is higher in non-epidemic state, so it can be concluded that the number of cases in the non-epidemic period has more fluctuations than the epidemic periods.

In addition, according to the simple MSM, the transition probabilities were calculated as follows:

$$p'' = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \begin{bmatrix} p_{00} & 1-p_{00} \\ 1-p_{11} & p_{11} \end{bmatrix}'' = \begin{bmatrix} 0.935 & 0.064 \\ 0.017 & 0.982 \end{bmatrix}$$

P_{00} is indication of the non-epidemic state. It is 93.5% likely to stay in non-epidemic state, and only 6.5% is likely to change from a non-epidemic to epidemic state. P_{10} indicates a 1.7% chance of regime change from epidemic to non-epidemic state. In other words, if the regime is in an epidemic state, it is 98.2% likely to stay in epidemic state.

The following equations were used to calculate the average period in each regime:

$$\text{The average length of the nonepidemic state period} = \frac{1}{p_{01}} = \frac{1}{0.064} = 15.64 \text{ months}$$

$$\text{The average length of the epidemic state period} = \frac{1}{p_{10}} = \frac{1}{0.017} = 58.82 \text{ months}$$

That is, when the regime is in the non-epidemic state, it lasts an average of 15 months, and when it enters the epidemic state, it lasts an average of 58 months, and then enters a non-epidemic state.

The multiple MSM shows that among the environmental variables

Table 4. Regression coefficients of simple and multiple MSM for cutaneous leishmaniasis cases in Isfahan province [Seasonality controlled with the parametric method of periodic function (Fourier terms)].

Parameters	Simple MSM				Multiple MSM			
	Coefficient	S.E	95% *CI*	*P*-value	Coefficient	S.E	95% *CI*	*P*-value
Sinus	-1.67	0.03	-1.74, -1.59	<0.001	-1.59	0.120	-1.830, -1.360	<0.001
Cosinus	0.50	0.03	0.42, 0.58	<0.001	0.73	0.093	0.550, 0.910	<0.001
Constant (0)	3.97	0.06	3.84, 4.10	<0.001	4.90	0.710	3.580, 6.390	<0.001
Constant (1)	4.68	0.03	4.62, 4.74	<0.001	7.05	0.900	5.270, 8.820	<0.001
Maximum relative humidity_lag 2 (0)	-	-	-	-	-0.003	0.003	-0.010, 0.003	0.297
Maximum relative humidity_lag 2 (1)	-	-	-	-	-0.005	0.003	-0.011, 0.001	0.114
Maximum relative humidity_lag 8 (0)	-	-	-	-	-0.001	0.003	-0.007, 0.004	0.737
Maximum relative humidity_lag 8 (1)	-	-	-	-	-0.008	0.003	-0.016, -0.007	0.032
Minimum relative humidity_lag 2 (0)	-	-	-	-	-0.002	0.013	-0.028, 0.023	0.835
Minimum relative humidity_lag 2 (1)	-	-	-	-	0.013	0.013	-0.013, 0.039	0.326
Maximum wind speed_lag 1 (0)	-	-	-	-	-0.014	0.016	-0.045, 0.172	0.377
Maximum wind speed_lag 1 (1)	-	-	-	-	-0.037	0.013	-0.064, -0.011	0.005
Maximum wind speed_lag 2 (0)	-	-	-	-	-0.0216	0.017	-0.056, 0.012	0.219
Maximum wind speed_lag 2 (1)	-	-	-	-	-0.033	0.015	-0.062, -0.003	0.029
Maximum wind speed _lag 3 (0)	-	-	-	-	-0.002	0.016	-0.033, 0.029	0.901
Maximum wind speed_lag 3 (1)	-	-	-	-	0.008	0.013	-0.018, 0.034	0.558
Minimum wind speed_lag 8 (0)	-	-	-	-	-0.263	0.078	-0.417, -0.108	0.001
Minimum wind speed_lag 8 (1)	-	-	-	-	-0.326	0.081	-0.486, -0.166	<0.001
Number of days with dust lag 7 (0)	-	-	-	-	-0.015	0.020	-0.056, 0.025	0.459
Number of days with dust lag 7 (1)	-	-	-	-	-0.028	0.016	-0.060, 0.004	0.089
Mean NDVI_lag9 (0)	-	-	-	-	3.64	2.05	-0.383, 7.670	0.076
Mean NDVI_lag9 (1)	-	-	-	-	0.737	2.55	-4.270, 5.750	0.773
Mean NDVI_lag11 (0)	-	-	-	-	3.34	2.25	-1.080, 7.760	0.139
Mean NDVI_lag11 (1)	-	-	-	-	3.63	2.46	-1.200, 8.470	0.141
Sigma	0.40	0.01	0.36, 0.44	-	0.310	0.015	0.281, 0.342	-
P_{00}	0.93	-	-	-	0.95	-	-	-
P_{10}	0.017	-	-	-	0.047	-	-	-
AIC	1.24	-	-	-	0.95	-	-	-

MSM, Markov switching model; S.E., standard error; *CI*, confidence interval; AIC, Akaike'information criterion; "-" not applicable.

included in the model, the variable of minimum wind speed at a lag of 8 months (P<0.001) has an effect on the non-epidemic state.

The variables of relative humidity maximum at a lag of 8 months (P=0.032), the minimum wind speed at a lag of 8 months (P<0.001), the maximum wind speed at a lag of 1 month (P=0.005), and the maximum wind speed at a lag of 2 months (P=0.029) were significantly associated with CL epidemics. The MSM estimated with the environmental variables was a better fit than the simple MSM model (without independent variables), with the lowest AIC value (Table 4).

Figure 5 shows the number of observed and predicted CL cases for the period between January 2000 and December 2019 based on the multiple MSM. The predicted model was completely fitted to the actual data (Figure 5). Also, the calculated MAPE was 3.5%.

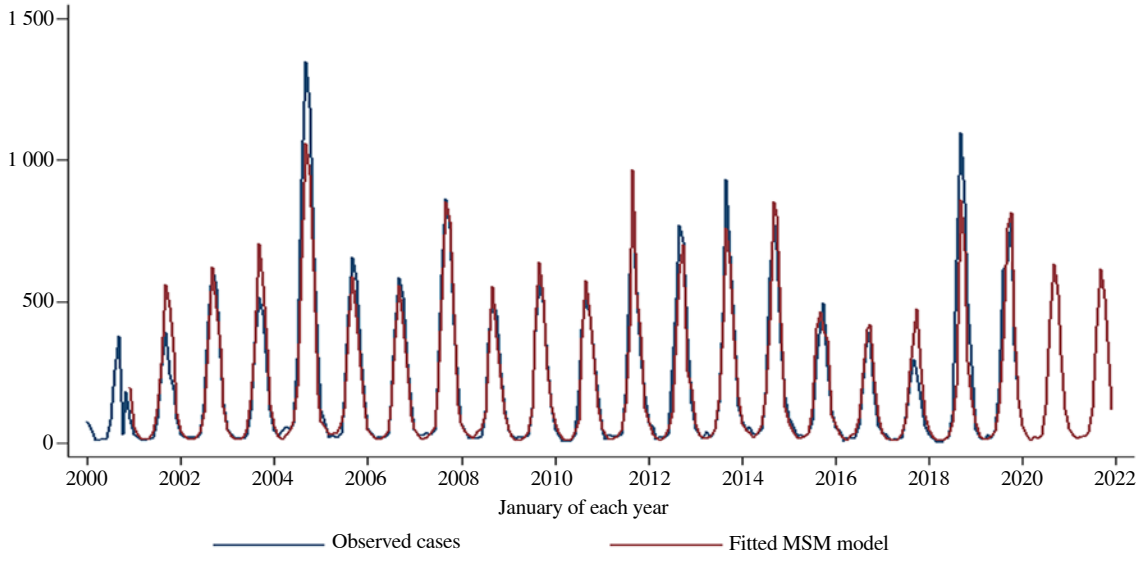

Figure 5. The observed and predicted cutaneous leishmaniasis cases between January 2000 and December 2019 based values on the multiple MSM.

3.5. Comparison of SARIMA and MSM models

Comparing both simple and multiple SARIMA models with simple and multiple MSM, it can be seen that AIC, BIC, and MAPE in MSM are much smaller than SARIMA models. Regarding the correlation between environmental variables in the SARIMA model, it was observed that there was significant association of CL cases with the variables of minimum relative humidity at a lag of 2 months and maximum wind speed at a lag of 3 months. While in the MSM, there was significant association of CL cases with the variables of maximum relative humidity at a lag of 8 months, minimum wind speed at a lag of 8 months, maximum wind speed at a lag of 1 month, and maximum wind speed at a lag of 2 months. However, the advantage of MSM is separating the effect of the independent variables individually in the epidemic and non-epidemic state, which in turn can provide the policymaker with the transition probabilities and the duration of each state.

4. Discussion

Time series analysis of surveillance data on infectious diseases is important to stimulate new hypotheses, predict observed events, and subsequently establish a quality control system[18,20]. The results of this study showed that the trend of CL is seasonal. The trend started to increase in July and reaches its peak in September and October and then starts to decline. This result can be justified considering the biology of vector (sandflies) and activity of reservoir (rodents) of Isfahan province[3,10]. Thus, in this province, the activity of *P. ansari*, *P. sergenti*, *P. caucasicus*, and *P. papatasi* as the main vectors of CL in Isfahan, continues from early April to August and the activity of *Rhombomys opimus* as the main reservoir is in spring and summer[3,21]. According to the extrinsic incubation period in sandflies (3 days) and incubation period in humans (approximately 4 months), it is compatible with the CL trend in this region[8,22].

Furthermore, our finding showed that the CL is in an epidemic state for 58 months, and a 15-months period of non-epidemic state with a recurrence of five years enters the epidemic stage. Therefore, it can be concluded that the interventions that have been made to control the disease for 20 years have not been very effective and the disease has gone through its normal course.

In the recent study which focused on Isfahan province between 2007 to 2015, there is no indication of CL incidence, and the incidence had been decreased by the end of the study[10]. This discrepancy in results can be attributed to the fact that this study considered the unit of time as year in eight time periods, while in the present study the period as month in 240 time periods (20 years) studied.

The results of this study are based on two reliable time series models that showed that minimum relative humidity, maximum relative humidity, minimum wind speed, and maximum wind speed were significantly associated with CL epidemics in different lags (P<0.05).

Numerous researches conducted in Iran have shown an association between environmental factors and CL occurrence. In a study conducted in Isfahan province by Ramezankhani *et al.* using negative binomial regression method, they showed that there is a significant association of mean temperature, relative humidity, land slope, maximum wind speed, rainfall, altitude, and vegetation with the incidence of CL[23]. Also, a similar study using the SARIMA time series model run in Fars province of Iran showed that the rainfall, relative humidity, and rainy days are associated with the incidence of CL[8]. Similar results were reported in Golestan Province, Iran[24]. There are reported studies in other parts of the world such as Tunisia[25], Tanzania[26], Algeria[27], and Afghanistan[28]. All of these found that there is a positive correlation between rainfall, humidity, and mean temperature with the incidence of CL. These effects can be explained due to the impact of these factors on the spread, survival, and activity of *Leishmania* parasite, its reservoirs, and vectors of this disease. Increasing the population of sandflies can increase the number of bites and consequently increase the transmission of infection. Increasing the minimum temperature and humidity reduces the maturation period and the extrinsic incubation period in sandflies, which can increase their population and consequently the rate of infection transmission[29,30]. In addition, female sandflies usually choose places with high humidity and temperature to lay their eggs so that their larvae can have the required humidity and temperature for their survival to provide easy feeding conditions[8,31]. The correlation between wind speed and some vector-borne diseases has been proven. Accordingly, wind has dual effects on the vectors and the hosts. A high-speed wind can reduce the possibility of mosquito bites; on the other hand, it can increase the flight length of sandflies[10,32].

In this study, we firstly performed pre-whitening to evaluate the relationships between the incidence of CL and the environmental data at different lags. Then, the cross-correlation between the residuals of SARIMA models in different series was investigated[8,15]. In time series analysis, cross-correlation on the original data is not recommended because without pre-whitening, the significant correlations observed between the one lags of different variables may be due only to the auto-correlation in seasonal time series[33]. In addition, the statistics such as Pearson and Spearman correlation coefficients assume that the data are independent, whereas this assumption is not true in time series data. Each observation is correlated with its previous values, so using such statistics without pre-whitening can result in a misleading outcome.

The relationship between environmental parameters and CL occurrence in different lags is attributed to the indirect occurrence of disease with sandfly activity at different times[29]. It is found that the CL is more common in autumn and winter, hence it is better to get more insight into this and how they are related. For example, with the noticeable change of environmental factors from a few months ago to the health system, there should be an alert about the epidemic of the disease, which in this study showed the models up to 8 months ago, *i.e.* at lag of 8 months.

In recent years, the use of ARIMA models in medicine and health, especially zoonotic diseases, has been increasing. Rahmanian *et al.* used the ARIMA (3,0,4) model to forecast human brocellosis in Yazd province, Iran[18]; Tohidinik *et al.* applied SARIMA (2,0,0) (2,1,0,12) models to forecast zoonotic CL in Fars province, Iran[8]. Yang et al. used ARIMA (0,2,1) to predict incidence of Malta fever in China[11] and Esmaeilzadeh *et al.* used ARIMA (1,0,0), ARIMA (1,0,1), and ARIMA (1,0,1) for determining the new confirmed cases, the death cases, and the recovery cases of COVID-19 in Iran[20].

The main idea of using Markov models in epidemiological surveillance of infectious diseases was established in 1999[34], but few studies have used the MSM to forecast zoonotic diseases, especially leishmaniasis. The results of this study showed that despite the relatively acceptable predictive power of SARIMA and MSM models, the latter provides more information to policymakers and is more helpful in planning for future interventions than the former model.

One of the strengths of this study is that the time unit considered is much longer than other time-series studies, *i.e.* 240 months, which reveals more accurate and detailed information about the trend of disease and its link with the environmental factors. Also, to our best knowledge, this is the first study that simultaneously uses the SARIMA and MSM of time series analysis for CL and compares them.

The limitations of this study lie in the fact that there are other parameters than the climatic factors such as host-related factors, vectors, reservoir diversity, parasite type, soil type, and health interventions performed in the epidemiology of CL in this region which are effective and should be considered for the future studies.

In conclusion, the results of this study showed that SARIMA (1,0,1) (1,1,1) 12 model and MSM can be a useful tool for predicting CL in Isfahan province. The time-series models can be used as a complementary tool of national disease surveillance systems for early warning of epidemics. This can help policymakers as a decision support tool in the preparedness of public health problems before the start of epidemic of the disease.

Since infectious diseases fall into one of two states of epidemic and non-epidemic, the use of MSM (dynamic) in this case is recommended as it provides more information to model as compared to single distribution (Box-Jenkins SARIMA) for all observations. The MSM seems to be suitable for the disease which does not have a seasonal trend and their regime is changed in the long term. Therefore, we recommend that in future studies, this model should be applied for diseases that do not have a seasonal trend.

Conflict of interest statement

The authors declare that there is no conflict of interest.

Acknowledgements

This study was derived from Ph.D. thesis from faculty of Veterinary Medicine, University of Tehran, Tehran, Iran. The authors would like to thank health deputy of Isfahan University of Medical Sciences for trying to control these infections and perform primary health care.

Authors' contributions

The initial idea of this study was by VR. SB and VR contributed to the design of the work, data collection and writing of drafting the article and implementing all comments from the reviewers. VR and AAH participated in the data analysis and interpretation. AAH and MB contributed to a critical revision of the article. All authors read and approved the version to be published.

References

[1] Rahmanian V, Rahmanian K, Sarikhani Y, Jahromi AS, Madani A. Epidemiology of cutaneous leishmaniasis, West South of Iran, 2006-2014. *J Res Med Dent Sci* 2018; **6**(2): 378-383.

[2] Alvar J, Vélez ID, Bern C, Herrero M, Desjeux P, Cano J, et al. Leishmaniasis worldwide and global estimates of its incidence. *PloS One* 2012; **7**(5): e35671.

[3] Karami M, Doudi M, Setorki M. Assessing epidemiology of cutaneous leishmaniasis in Isfahan, Iran. *J Arthropod Borne Dis* 2013; **50**(1): 30-37.

[4] Holakouie-Naieni K, Mostafavi E, Boloorani AD, Mohebali M, Pakzad R. Spatial modeling of cutaneous leishmaniasis in Iran from 1983 to 2013. *Acta Trop* 2017; **166**(1): 67-73.

[5] Gholamrezaei M, Mohebali M, Hanafi-Bojd AA, Sedaghat MM, Shirzadi MR. Ecological Niche Modeling of main reservoir hosts of zoonotic cutaneous leishmaniasis in Iran. *Acta Trop* 2016; **160**(3): 44-52.

[6] Yaghoobi-Ershadi MR. Control of phlebotomine sand flies in Iran: A review article. *J Arthropod Borne Dis* 2016; **10**(4): 429-444.

[7] Norouzinezhad F, Ghaffari F, Norouzinejad A, Kaveh F, Gouya MM. Cutaneous leishmaniasis in Iran: Results from an epidemiological study in urban and rural provinces. *Asian Pac J Trop Biomed* 2016; **6**(7): 614-619.

[8] Tohidinik HR, Mohebali M, Mansournia MA, Niakan Kalhori SR, Ali-Akbarpour M, Yazdani K. Forecasting zoonotic cutaneous leishmaniasis using meteorological factors in eastern Fars province, Iran: A SARIMA analysis. *Trop Med Int Health* 2018; **23**(8): 860-869.

[9] Nikonahad A, Khorshidi A, Ghaffari HR, Aval HE, Miri M, Amarloei A, et al. A time series analysis of environmental and metrological factors impact on cutaneous leishmaniasis incidence in an endemic area of Dehloran, Iran. *Environ Sci Pollut Res Int* 2017; **24**(16): 14117-14123.

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

[11]Yang L, Bi ZW, Kou ZQ, Li XJ, Zhang M, Wang M, et al. Time-series analysis on human brucellosis during 2004-2013 in Shandong Province, China. *Zoonoses Public Health* 2015; **62**(3): 228-235.

[12]Chen S, Zhang H, Liu X, Wang W, Hou S, Li T, et al. Increasing threat of brucellosis to low-risk persons in urban settings, China. *Emerg Infect Dis* 2014; **20**(1): 126-130.

[13]Zhang X, Liu Y, Yang M, Zhang T, Young AA, Li X. Comparative study of four time series methods in forecasting typhoid fever incidence in China. *PLoS One* 2013; **8**(5): e63116.

[14]Lu HM, Zeng D, Chen H. Prospective infectious disease outbreak detection using Markov switching models. *IEEE Trans Knowl Data Eng* 2009; **22**(4): 565-577.

[15]Ansari H, Mansournia M, Izadi S, Zeinali M, Mahmoodi M, Holakouie-Naieni K. Predicting CCHF incidence and its related factors using time-series analysis in the southeast of Iran: Comparison of SARIMA and Markov switching models. *Epidemiol Infect* 2015; **143**(4): 839-850.

[16]Yan J, Wang L. Suitability evaluation for products generation from multisource remote sensing data. *Remote Sensing* 2016; **8**(12): 995.

[17]Iranian Space Agency. *Normalized difference vegetation Index (NDVI).* [Online]. Available from: http://www.isa.ir/fa/page/101227.html. [Accessed on 2 May 2020].

[18]Rahmanian V, Bokaie S, Rahmanian K, Hosseini S, Firouzeh A. Analysis of temporal trends of human brucellosis between 2013 and 2018 in Yazd Province, Iran to predict future trends in incidence: A time-series study using ARIMA model. *Asian Pac J Trop Med* 2020; **13**(6): 272-277.

[19]Muangkhoua S. Time series forecasting by using Box-Jenkins method. *Vajira Medical Journal: J Urban Med* 2019; **63**: S185-S192.

[20]Esmaeilzadeh N, Shakeri M, Esmaeilzadeh M, Rahmanian V. ARIMA models forecasting the SARS-COV-2 in the Islamic Republic of Iran. *Asian Pac J Trop Med* 2020; **13**(11): 521-524.

[21]Veysi A, Vatandoost H, Arandian MH, Jafari R, Yaghoobi-Ershadi MR, Rassi Y, et al. Laboratory evaluation of a rodenticide-insecticide, Coumavec®, against *Rhombomys opimus*, the main reservoir host of zoonotic cutaneouse leishmaniasis in Iran. *J Arthropod Borne Dis* 2013; **7**(2): 188-193.

[22]Falcão de Oliveira E, Oshiro ET, Fernandes WdS, Murat PG, de Medeiros MJ, Souza AI, et al. Experimental infection and transmission of *Leishmania* by *Lutzomyia cruzi* (Diptera: Psychodidae): Aspects of the ecology of parasite-vector interactions. *PLoS Negl Trop Dis* 2017; **11**(2): e0005401.

[23]Ramezankhani R, Sajjadi N, Nezakati Esmaeilzadeh R, Jozi SA, Shirzadi MR. Climate and environmental factors affecting the incidence of cutaneous leishmaniasis in Isfahan, Iran. *Environ Sci Pollut Res Int* 2018; **25**(12): 11516-11526.

[24]Mollalo A, Alimohammadi A, Shirzadi MR, Malek MR. Geographic information system-based analysis of the spatial and spatio-temporal distribution of zoonotic cutaneous leishmaniasis in Golestan Province, north-east of Iran. *Zoonoses Public Health* 2015; **62**(1): 18-28.

[25]Toumi A, Chlif S, Bettaieb J, Alaya NB, Boukthir A, Ahmadi ZE, et al. Temporal dynamics and impact of climate factors on the incidence of zoonotic cutaneous leishmaniasis in central Tunisia. *PLoS Negl Trop Dis* 2012; **6**(5): e1633.

[26]Talmoudi K, Bellali H, Ben-Alaya N, Saez M, Malouche D, Chahed MK. Modeling zoonotic cutaneous leishmaniasis incidence in central Tunisia from 2009-2015: Forecasting models using climate variables as predictors. *PLoS Negl Trop Dis* 2017; **11**(8): e0005844.

[27]Selmane S. Dynamic relationship between climate factors and the incidence of cutaneous leishmaniasis in Biskra Province in Algeria. *Ann Saudi Med* 2015; **35**(6): 445-449.

[28]Adegboye O, Al-Saghir M, Leung DH. Joint spatial time-series epidemiological analysis of malaria and cutaneous leishmaniasis infection. *Epidemiol Infect* 2017; **145**(4): 685-700.

[29]Abdolahnejad A, Mousavi SH, Sofizadeh A, Jafari N, Shiravand B. Climate change and distribution of zoonotic cutaneous leishmaniasis (ZCL) reservoir and vector species in central Iran. *Model Earth Syst Environ* 2020; **3**: 1-11.

[30]Trájer AJ. The potential impact of climate change on the seasonality of *Phlebotomus neglectus*, the vector of visceral leishmaniasis in the East Mediterranean region. *Int J Environ Health Res* 2019; **2**: 1-19.

[31]Erguler K, Pontiki I, Zittis G, Proestos Y, Christodoulou V, Tsirigotakis N, et al. A climate-driven and field data-assimilated population dynamics model of sand flies. *Sci Rep* 2019; **9**(1): 1-15.

[32]Wu X, Lu Y, Zhou S, Chen L, Xu B. Impact of climate change on human infectious diseases: Empirical evidence and human adaptation. *Environ Int* 2016; **86**: 14-23.

[33]McDowall D, McCleary R, Bartos BJ. *Interrupted time series analysis.* Oxford: Oxford University Press; 2019.

[34]Le Strat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. *Stat Med* 1999; **18**(24): 3463-3478.