Comparative Performance Analysis for Generalized Additive and Generalized Linear Modeling in Epidemiology

Methods of Evaluation for Modeling Disease Incidence

Talmoudi Khouloud [1,2,3,4], Bellali Hedia [3,4,5], Ben-Alaya Nissaf [5,6], Saez Marc [7], Malouche Dhafer [2], Chahed Mohamed Kouni [3,4,5]

[1] National Engineering School of Tunis, ENIT, Tunis El Manar University, Tunis, Tunisia.
[2] Research Unit on Modeling, Statistics and Economic Analysis (MASE, ESSAI), High School of Statistics and Information Analysis (ESSAI), University of Carthage, Tunis, Tunisia.
[3] Department of Epidemiology and Statistics, Abderrahman Mami Hospital, Ariana, Tunisia.
[4] Research Unit "Analysis of the Effects of Environmental and Climate Change on Health", Department of Epidemiology and Statistics, Abderrahmen Mami Hospital, Ariana, Tunisia.
[5] Department of Epidemiology and Public Health, Faculty of Medicine of Tunis, Tunis El Manar University, Tunis, Tunisia.
[6] National Observatory of New and Emergent Diseases, Tunis, Tunisia.
[7] Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Girona, Spain.

***Abstract*—Most environmental-epidemiological researches emphasize modeling as the causal link of different events (e.g., hospital admission, death, disease emergency). There has been a particular concern in the use of the Generalized Linear Models (GLMs) in the field of epidemiology. However, recent studies in this field highlighted the use of non-parametric techniques, especially the Generalized Additive Models (GAMs). The aim of this work is to compare performance of both methods in the field of epidemiology. Comparison is done in terms of sharpening the relation between the predictors and the response variable as well as in predicting outbreaks. The most suitable method is then adopted to elucidate the impact of bioclimatic factors on the emergence of the zoonotic cutaneous leishmaniasis (ZCL) disease in Central Tunisia. Monthly epidemiologic and bioclimatic data from July 2009 to June 2016 are used in this study. Akaike information criterion, R-squared and F-statistic are used to compare model performance, while the root mean square error is used for checking predictive accuracy for both models. Our results show the potential of GAM model to provide a better assessment of the nonlinear relations and to give a high predictive accuracy compared to GLMs. The results also stress the inaccurate estimation of risk factors when linear trends are used to model nonlinear structured data.**

***Keywords*—Generalized linear model; generalized additive model; zoonotic cutaneous leishmaniasis; Central Tunisia**

I. Introduction

In the last decade, there has been an increasing interest for the use of nonparametric modeling techniques in the field of epidemiology, especially the generalized additive models (GAMs) [1]. In fact, they became of great concern for modern analytics in several fields of scientific researches. However, researchers are still faithful to the use of parametric techniques such as the generalized linear models (GLM) [2]. This can be explained by their robustness and the reliability of the results provided. Since their emergence, both approaches have been extensively applied in diverse domain such as environment, signal processing, ecology and particularly in epidemiology [3]-[5]. This can be explained by their ability to describe the real dynamics existing in the data and to the straightforward way of interpreting and representing the results using graphical ways.

In fact, GLMs are parametric models that allow for non-linearity through the use of high order polynomial. The specification of these techniques over the traditional linear models is that the mean of the dependent variable is expressed by a linear combination of the independent variables through a link function. Also, the GLMs offer the possibility of using different families of probability distributions for the data in order to whiten the error structures and thereby allow a better fitting for complex relationships between a response and a set of independent variables.

The GAMs were first implemented by Hastie and Tibshirani [6] based on the backfitting algorithm. In this case, GAMs are purely nonparametric. However, to give more flexibility to the GAM approach, the alternative approach of the backfitting algorithm was developed and was based on penalized likelihood estimation. In this way, GAMs are considered as semiparametric method and offered the ability of including parametric as well as nonparametric terms. GAMs [7] are a smooth extension of the GLMs and thus, inherit from them the flexibility in modeling complex shapes, the use of various family distributions and the link functions.

Previous studies that have evaluated the performance of both statistical methods for determining the causal link between a set of explanatory variables on a response variable were based on simulated data or only limited for comparisons without interpreting results [8], [9]. Although GAMs are theoretically flexible on modeling tasks when compared to

GLMs, few studies give evidence conclusions based on real data.

In this paper, a case study of zoonotic cutaneous leishmaniasis (ZCL) is used as an illustrative example for environmental-epidemiological researches. In fact, ZCL is a neglected tropical disease and is considered as a public health concern in the regions of the Maghreb countries and the Eastern part of the Mediterranean regions, including Tunisia [10]. Infected female sandflies transmit the disease to humans living in rural areas, where pollution and unawareness of people are highly present. The life cycle of the ZCL is highly related to bioclimatic and environment change [11]. In Tunisia, the disease first emerged in 1982 in the region of Sidi Bouzid, central Tunisia [12]. Recent studies showed that the epidemic of ZCL occurs every 4 to 7 years in that region [13], [14].

Few works studied the modeling of the impact of climate factors on transmission of leishmaniasis infection. Talmoudi et al. [4] used the generalized additive (mixed) models to assess the relationship between environmental and bioclimatic factors and ZCL occurrence in central Tunisia. They found that rainfall, temperature, relative humidity and rodent's density are the main factors influencing the incidence of the disease. Toumi et al. [14] used GAM and generalized estimating equations to give seasonal distributions of the ZCL disease in central Tunisia and to determine the relative importance of significant factors on the disease. They found that temperature and humidity are the main predictors of the ZCL incidence. Moreover, few studies give descriptive patterns of the temporal distributions of vector-borne diseases emergence using the GLM [15], [16]. However, there were no study modeling such relation in the case of ZCL and using the GLMs.

In this study, the performance of GLM and GAM was first compared in terms of determining the impact of bioclimatic variables on the emergence of ZCL disease and second, in their predictive accuracies.

This paper is organized as follows: in Section 2 the context of discussing the way to use statistical models in the scope of epidemiological studies is detailed. In Section 3, principle findings obtained from comparison of models in terms of performances and from predictive accuracy are shown through real data.

II. Use of Statistical Models in Epidemiological Studies

The advantage of statistical methods is the ability to provide a mathematical equation that reflects the complexity of the relation between a response variable and a set of explanatory variables. Moreover, plausible reasons for their use include the availability of easy tools for interpreting results, checking for the significant predictors, assessing their relative importance and their display graphics. In epidemiology, this type of model may better explain the real dynamics of the phenomena.

The use of GLMs and GAMs in our study meets our needs in explanatory analysis and predictive challenging tasks. In fact, on one hand, explanatory analysis seeks to characterize relevant factors that impact the ZCL emergency. On the other hand, predictive models are used to help government policy makers take right decisions to reduce the spread of the disease.

A. Generalized Linear Models

In the generalized linear model (GLM), the response variable, y_i, depends on a smooth monotonic function of the linear predictor through a link function g and thereby, allow for non-linearity and non-constant variance structure in the data [2]. The response variable is assumed to have normal distribution or other distributions from the exponential family. It is modeled as the sum of linear predictors η_i and a random error term with zero mean. The response variable is considered as a weighted sum of p predictor variables, x_j , with an intercept, β_0 , and Gaussian error with standard deviation σ. Then, the generalized linear model is described as:

$$g(E(y_i)) = \eta_i + \epsilon_i, where\ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_i x_{ji}$$

In our case the dependent variable, y_i , is the reported number of ZCL cases and x_j accounts for the bioclimatic and environmental variables. In this work, the R package version 3.3.3 [17] is used to conduct all statistical analysis. The GLM models were constructed using the *glmulti* [18] R package.

B. Generalized Additive Models

The GAMs are semi-parametric approaches that allow for the inclusion of distribution network of independent variables as a function of the response variable [7], [19]. The response variable y_i is modeled by a sum of smooth functions of covariates. The general structure of GAM can be written as:

$$\begin{cases} g\big(E(y_{it})\big) = \theta X_{it} + \sum_{j=1}^{p} f_j(x_i) \\ y_{it} = E(y_{it}) + \epsilon_{it} \end{cases}$$

Where g is a link function, y denotes the response variable, the vector θ contains fixed parameters, X_{it} is a row of fixed effects matrix, and f_j are smoothing splines of the p explanatory covariates, x_i. The residual errors ϵ_i are random Gaussian noise with mean 0.

The smooth functions in GAMs can be defined by a piecewise polynomial functions or basis functions. The locations where polynomials are connected are called knots, denoted by k. The most common basis functions used are thin plate regression splines and cubic regression splines [20]. In the case of environmental epidemiological studies where climate factors are the predictors, the cubic spline functions are the most adopted. In these splines, if b_k is the kth cubic basis function, then f is represented as:

$$f(x) = \sum_{k=1}^{K} b_k(x)\beta_k$$

Where K is the total number of knots and β_k are unknown parameters. Thus, by using the cubic splines the curve obtained

is a gathering of cubic polynomials which are continuous in value and in first and second derivatives.

Smooth functions are estimated by adding a "wiggliness" penalty to penalized least squares [17]. In GAMs, the dilemma is to minimize to following equation:

$$\|y - X\beta\|^2 + \lambda \int_0^1 [f"(x)]^2 dx$$

Where λ is a smoothing parameter. The first statement measures the adequacy of the function to data. The integral of squared second derivatives is used to penalize too "wiggly" models and to control the degree of smoothness (edf) given for curvature in functions. If the edf is equal to 1, then the relation is estimated to be linear. An edf of 3 indicates a quadratic shape. The higher the edf, the more wiggly the relation.

Because f is linear in the parameters β_i, the penalty can be written as a quadratic form in β:

$$\int_0^1 [f"(x)]^2 dx = \beta^T S \beta$$

Where S is a matrix ok known coefficients.

Therefore, to avoid overfitting, the smoothing parameters λ are controlled by minimizing the Generalized Cross Validation (GCV) score [21]. It can be summarized by:

$$\mathcal{V}_g = \frac{n \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{[tr(I - A)]^2}$$

Where A is an influence matrix or a hat matrix and can be written as: $A = X(X^T X + \lambda S)^{-1}$. The GCV score have computational advantages in terms of invariance.

The use of GAM in the epidemiological studies is helpful, especially when the relation between the response and the explanatory variable is not known in advance. The mgcv R package [7] is used in this work to build the GAM models.

C. Model Selection and Validation

In this study, where the regression techniques are used, it is important to assess the adequacy of the model by testing the properties of the residual error. Specific tests are used to check if the residuals are conform to three major assumptions with each model to ensure optimum utility of the models: (1) first, test for non-correlated or random errors. For example, in our case where time series data are used, temporal correlation is likely to be present. But when this step is ignored, there is tendency of misspecification of estimators. (2) Second, check for homoscedasticity for error variance (constant variance). (3) The third assumption is that errors are normally distributed with a mean 0.

Next, in order to evaluate performance of the models, a cross-validation test is conducted. A training set containing 80% of the data is selected to build the model. This training sample is used to make the validation for the remaining 20% of the data (test sample). This sub-sampling process was repeated several times for each cross-validation for both GLM and GAM models. The cv.gam and cv.lm functions in gamclass and DAAG packages under R software are used, respectively. The same steps for verification of goodness-of-fit for GLM and GAM models were followed through performing plots of residuals and standard errors as well as model validation using cross-validation test.

Third, in order to assess the prediction performance of the models, the last 20% of the data were used to evaluate offset between predicted values of the model and the original values. This can be checked based on the root mean square error (RMSE), which is a reliable measure [22]. It is calculated as the root of sum squared difference between predicted values of a model and original values, divided by the size of the data. The model with smaller RMSE has a better predictive performance.

D. Performance Comparison

Models are compared by means of the Akaike Information Criterion (AIC), the R-squared statistic and F-statistic. The AIC [23] measures the goodness of fit and is given by $= -2\ln(l) + 2P$, where l is the likelihood and P is the number of parameters in the model. The AIC score can be calculated for both parametric and non-parametric models. The lower the AIC value, the better-fit the model is for the data. Besides, the coefficient of determination, the R-squared value indicates the variability of the response data explained by the model. A higher R-squared lends greater explanatory power to this model. In addition, the F-statistic is a measure adopted to compute the regression strength. It is given by $F = \left(\frac{SS_1 - SS_2}{df_1 - df_2}\right) / \left(\frac{SS_2}{df_2}\right)$, where SS_1 and SS_2 are the sum of squares for model 1 and model 2, respectively. While df_1 and df_2 are the degrees of freedom for each model. The higher the F-statistic, the better-fit the model is for the data.

Moreover, to examine the adequacy of the models, an autocorrelation function (ACF) and a partial autocorrelation function (PACF) plots were used to check for independent and random distribution of the residuals over time from the established models.

III. Results

A. Data Collection and Exploratory Analysis

Monthly data of ZCL cases were collected between July 2009 and June 2016 in three rural regions of Sidi Bouzid, central Tunisia. Bioclimatic factors were obtained from an active surveillance system implemented in this region and include: average temperature in degree Celsius, cumulative rainfall in millimeters, relative humidity in percentage and rodent density estimated according to their activity. We used data from July 2009 to June 2015 (72 months) for training set and data from July 2015 to June 2016 (12 months) as test set.

A total of 2125 ZCL counts were reported during the study period from 2009 to 2015. There was a sharp increase of ZCL cases during the cold season for each year (Fig. 1), in particular for months from September to January. The observations of ZCL were significantly autocorrelated as reported by the autocorrelation function plot (Fig. 2). An outbreak of 387 ZCL case was noticed between August 2013 to January 2014 (the fifth epidemic season). Also, a significant peak of the disease is seen between August 2015 to January 2016 (about 51.8% of all notified ZCL cases). Therefore, a moving average of order 2

was run on the original data for adjusting the estimation of a monthly ZCL incidence distribution. This adjustment for the response variable is presented in Fig. 3.

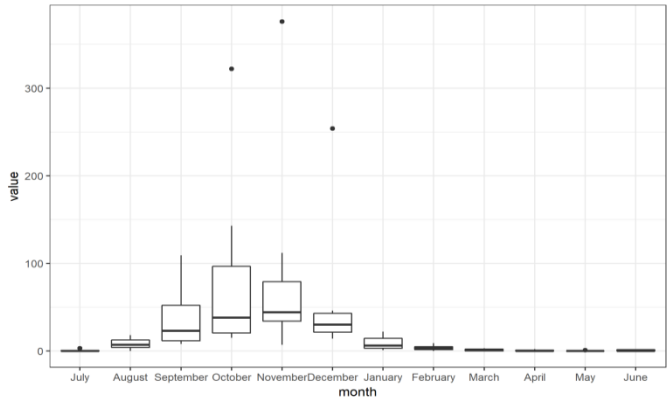

Fig. 1. Box plot for monthly ZCL incidence 2009-2016.

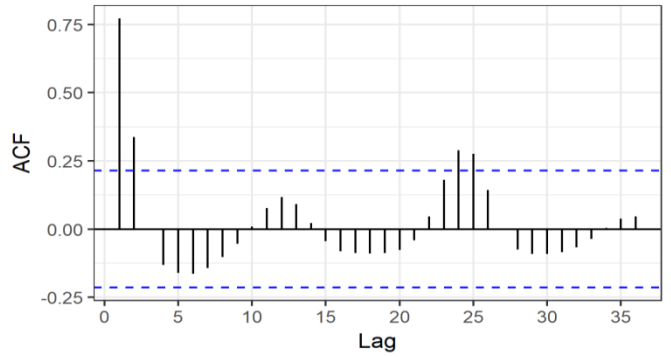

Fig. 2. Correlation between observations of the response.

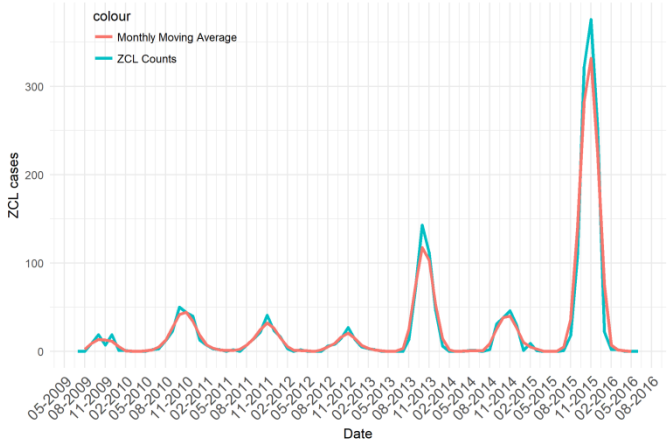

Fig. 3. Month and year of ZCL lesion onset.

B. Model Selection

Results for comparing performances are shown in Table I. A first comparison of the GLM and GAM models was performed based on the lowest AIC likelihood ratio. Regarding the AIC criterion, results showed that GAM model has the lower score compared to GLM (AIC-GLM = 389.10; AIC-GAM = 382.44). It indicates a high quality for model performance for GAM. In addition, results from R-squared and F-statistics stressed the results obtained from AIC score. In fact, a GAM model explains more variability on the data (69%) compared to GLM (45%). Results again showed the outperformance of GAM over the GLM in indicating the goodness of fit (Table I).

Fig. 4 showed the dissimilarities between two models concerning the residual variances. The GLM model had more dispersed residuals compared to GAM and was discarded as possible candidate model. GAM achieved better results concerning residual variances and presented the best overall performance. However, the ACF and PACF plots for both models showed that the errors were whitened and validated the assumption of residuals normality.

TABLE I. AIC, R-SQUARED AND F-STATISTIC FROM GLM AND GAM MODELS

Model	AIC	R-squared	F-statistic
GLM	389.10	0.45	1.11
GAM	382.44	0.69	2.56

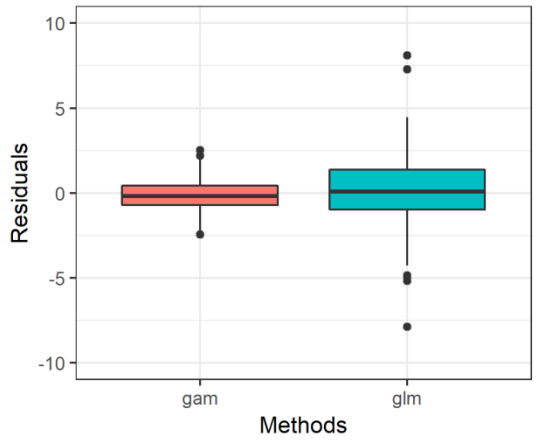

Fig. 4. Residuals for each modeling technique.

C. Results from the Selected Model

Fig. 5 showed the fitted functions of the GAM model for ZCL incidence in relation to each climate factor. Indeed, the average temperature delayed 4 months is highly associated with the emergence of the disease in cold seasons (Fig. 5(A)). This association is very wiggly (edf = 5.86). An overall increase of number of ZCL cases is reported if average temperature is between 5°C and 23°C. A decrease effect for ZCL incidence is seen when temperatures are higher than 23°C. The relation shape of rodent density is illustrated in Fig. 5(B). A density around 30 can be a risk factor of emergence for the disease.

The impact of rainfall on ZCL incidence (Fig. 5(C)) is very wiggly (edf = 7.61). It represents a lot of fluctuations, indicating the sensitivity of the transmission of the disease according to the value of the cumulative rainfall. Moreover, an increasing association is observed between the number of cases and the relative humidity delayed 5 months under 50% (edf = 2.34, $p < 0.001$). A decreasing effect is shown for values of relative humidity that are higher than 50% (Fig. 5(D)).

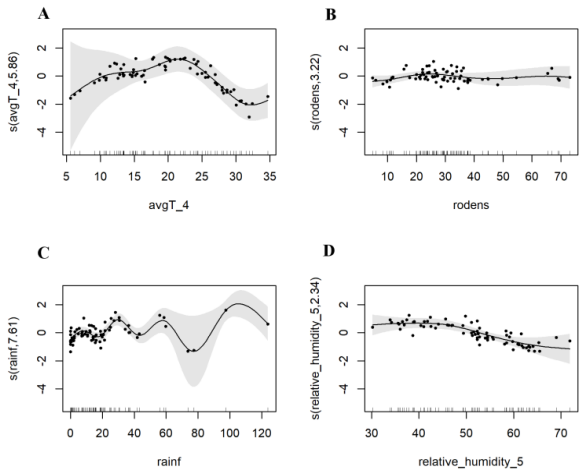

Fig. 5. Relattioship between bioclimatic factors and ZCL emergnece using the best fit GAM model.

D. Prediction

The GAM model has shown its effectiveness in predicting the number of cases. On one hand, the agreement of predicted and original values is seen with the high correlation coefficient (cor = 0.81, confidence interval = [0.45 - 0.94], p = 0.001 < 5%). On the other hand, the prediction from GAM has reported the same appearance as the original values. That is, a significant increase in the number of cases between October and January and a low incidence during the warm months are seen during the last epidemic season (Fig. 6). However, the prediction of the GLM model is far from the original values. Indeed, the correlation between the fitted values from GLM and the original values is very weak and not significant (cor = 0.06, p = 0.84 > 5%). Also, the confidence bands for GLM are very large and the seasonality of the disease in not detected. In addition, the RMSE from the GLM is higher than the one from the GAM model (RMSE-GLM = 301.18, RMSE-GAM = 285.95). This supports the fact that the precision of prediction with the GAM model is better than with the GLM.

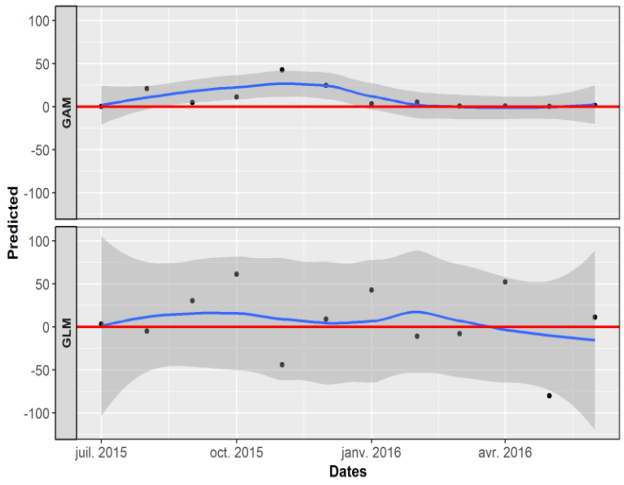

Fig. 6. Prediction from both GLM and GAM models.

IV. Conclusion

In this paper, the challenging task of focusing on the performances of the GLMs and the GAMs in determining relevant bioclimatic factors responsible for the emergence of the ZCL disease was investigated. The use of these models was advocated and recommended when modeling complex relations in epidemiology [24].

Results from the semi-parametric GAM models proved to be efficient to depict the nonlinear effects of independent variables on the outcome. According to Wood [19], the advantage of this method is its adaptation with the nonlinear shapes and the potential interaction effects through possible use of different smooth functions. This modeling framework allows researchers to explore the effect of explanatory variables in a flexible way than allowed under the GLMs or traditional methods. The results of this study provided important insights into the relative strength of the GAM methodology in the field of environmental-epidemiological studies. This stressed the fact that modeling such relation need for robust techniques to assess real dynamics.

Regarding the GLMs, they provide less performances than the GAMs in this study. This is due to the complex and undetermined relation existing in our case study. However, GLMs can provide robust conclusions in epidemiological studies if the relation between independent variables and the response variable is known in advance.

GAMs can be also adopted to construct a predictive model for vector borne diseases. In fact, GAM models showed better fit and good prediction accuracy when compared to GLMs. The competitively of GAMs in prediction task was highlighted in previous works [25].

The contribution of our study consists of the recommendation of the use of GAM models in the field of epidemiology where causal link need to be assessed. The practical utility of this method was demonstrated through a real data analysis. We believe that this methodology may help policy makers in evaluating the impact of risk factors on the transmission of the ZCL and thus, help to reduce the spread of this disease. Also, the GAM model can be adopted in any epidemiological study where the impact of climatic factors on the incidence of a disease needs to be resolved.

One limitation of our work is the availability of a short term series in this field. But, in order to validate the appropriate model and the significant predictors, a large dataset, over at least 10 years, is required.

The results obtained in this study encourage further explorations for the use of novel statistical techniques in the field of epidemiology. We would like to study the issues of deep learning methods, such as multilayer neural network or reinforcement learning. These techniques will help biostatisticians to identify relevant variables, particularly when complex relations need to be assessed.

REFERENCES

[1] T. Hastie, R. Tibshirani, "Generalized additive models," Stat Sci. 1986, 1(3): 297-318.

[2] P. McCullagh and J. A. Nelder, Generalized Linear Models. Monographs on Statistics and Applied Probability. London: Chapman and Hall, 1989.

[3] Y. L. Cheong, K. Burkart, P.J. Leitão, T. Lakes, " Assessing weather effects on dengue disease in Malaysia," Int. J. Environ. Res. Public Health, 2013, 10, pp. 6319-6334.

[4] K. Talmoudi, H. Bellali, N. Ben-Alaya, M. Saez, D. Malouche, M. K. Chahed ,"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.

[5] R. W. Shadish, A. F. Zuur, K. J. Sullivan, "Using generalized additive (mixed) models to analyze single case designs," Journal of School Psychology, 2014, 52: 149-178.

[6] T. Hastie, R. Tibshirani, Generalized Additive Models. London: Chapman & Hall, 1990.

[7] S. N. Wood, Generalized Additive Models: An Introduction with R. New York: Chapman & Hall/CRC, 2006.

[8] A. Guisan, Jr; T. C. Edwards, and T. Hastie, "Generalized linear and generalized additive models in studies of species distributions: setting the scene," Ecological Modelling, 2002, 157: 89-100.

[9] M. Austin, "Species distribution models and ecological theory: A critical assessment and some possible new approaches," Ecological Modelling, 2007, 200: 1-19.

[10] World Health Organization, 2010. Control of the leishmaniases. Report of a meeting of the WHO Expert Committee on the Control of Leishmaniases, Geneva, 22-26 of March.

[11] H. Bellali, N. Ben-Alaya, Ahmadi Z, Ennigrou S, Chahed MK, "Eco-environmental, living conditions and farming issues linked to zoonotic cutaneous leishmaniasis transmission in Central Tunisia: a population basedsurvey," Int J Trop Med Public Health, 2015, 5(1): 1-7.

[12] R. Ben Ismail, L. Gradoni, M. Gramiccia, S. Bettini, M. S. Ben Rachid, A. Garraoui, "Epidemic cutaneous leishmaniasis in Tunisia: Biochemical characterization of parasites," Trans R Soc Trop Med Hyg, 1986, 80: 669-70.

[13] H. Bellali, K. Talmoudi, N. Ben Alaya, M. Mahfoudhi, S. Ennigrou, M. K. Chahed, "Effect of temperature, rainfall and relative density of rodent reservoir hosts on zoonotic cutaneous leishmaniasis incidence in Central Tunisia," Asian Pac J Trop Dis, 2017, 7(2): 88-96.

[14] A. Toumi, S. Chlif, J. Bettaieb, N. Ben Alaya, A. Boukthir, Z. E. Ahmadi, A. Ben Salah, "Temporal dynamics and impact of climate factors on the incidence of zoonotic cutaneous leishmaniasis in Central Tunisia," PLoS Negl Trop Dis, 2012, 6: e1633.

[15] I. Alkhaldy, "Modelling the association of dengue fever cases with temperature and relative humidity in Jeddah, Saudi Arabia-A generalized linear model with break-point analysis," Acta Trop, 2017, 168: 9-15.

[16] L. G. L. Jumi, "Generalized linear models of Malaria incidence in Jubek State, South Sudan," Science Journal of Applied Mathematics and Statistics, 2017; 5(4): 134-138.

[17] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, V., Austria. 2015.

[18] V. Calcagno and C. de Mazancourt, "glmulti: An R package for easy automated model selection with (Generalized) linear models," Journal of statistical software, 2010, volume 34-N12.

[19] S. N. Wood, Generalized Additive Models: An Introduction with R, 2nd edition. Chapman & Hall/CRC press. Taylor & Francis, 2017.

[20] S. N. Wood, N. Pya, B. Säfken, "Smoothing Parameter and Model Selection for General Smooth Models," Journal of the American Statistical Association, 2016; 111(516): 1548-1575.

[21] S. N. Wood, "Modelling and smoothing parameter estimation with multiple quadratic penalties," Journal of the Royal Statistical Society B , 2000, 62: 413–428.

[22] R. J. Hyndman, A. B. Koehler, "Another look at measures of forecast accuracy," International journal of forecasting, 2006, 22(4): 679-688.

[23] H. Bozdogan, "Akaike's information criterionand recent developments in information complexity," Journal of Mathematical Psychology, 2000, 44: 62–91.

[24] T. Hastie, R. Tibshirani, "Generalized additive models for medical research," Stat Methods Med Res, 1995, 4(3): 187-96.

[25] M. Marmion, J. Hjort, W. Thuiller, M. Luoto, "A comparison of predictive methods in modelling the distribution of periglacial landforms in Finnish Lapland," Earth Surf Process and Landforms, 2008, 33: 2241-2254