# APPLICATION DES MODELES DE PREVISION UNIVARIEE ET MULTIVARIEE POUR LA LEISHMANIOSE CUTANEE ZOONOTIQUE DANS LE CENTRE DE LA TUNISIE

Amine Toumi[1] & Dhafer Malouche[1] & Afif BenSalah[2]

[1] U2S,ENIT Tunisie toumiamine@gmail.com

[1] U2S, ENIT Tunisie dhafer.malouche@me.com

[2] Laboratoire d'épidémiologie médical, IPT afif.bensalah@pasteur.rns.tn

**Résumé.** La leishmaniose cutanée causée par Leishmania major constitue un véritable problème de santé publique dans les pays du Maghreb et l'Est de la région méditerranée. Par conséquent, la prévision des épidémies et l'alerte précoce reste une priorité de la recherche pour les programmes de contrôle dans le monde ancien de leishmaniose cutanée en l'absence d'un vaccin efficace. Toutefois, pour cette menace de santé, ces systèmes ne peuvent pas être correctement conçus en l'absence d'évaluation fiable et valide de l'évolution temporelle des épidémies et de leurs déterminants. L'objectif du présent projet est d'analyser les tendances temporelles de la leishmaniose à *L.major* à Sidi Bouzid en utilisant des données mensuelles (1991-2007) et d'évaluer l'impact des variables climatiques sur le risque de la maladie. Plus spécifiquement, cette étude vise à établir un modèle de prévision de la LCZ et d'évaluer l'impact relatif des facteurs climatiques sur le risque de la LCZ. La technique de prévision basée sur la méthode de Box et Jenkins a montré que l'historique des observations ainsi l'aléa décalés par un seul mois peuvent prévoir quasi- parfaitement l'incidence de la LCZ. La présente étude a confirmé la relation entre la leishmaniose cutanée de l'ancien monde et les facteurs climatiques grâce aux modèles autorégressifs multivariés.

**Mots-clés.** ARIMA, VAR, AR multivariée, LCZ, Climatique

**Abstract.** Old world Zoonotic Cutaneous Leishmaniasis (ZCL) is a vector born disease caused by Leishmania major and transmitted by sand flies. It is endemic in the Middle East and North Africa. Like other vector born diseases, ZCL is highly sensitive to environmental and climate factors. However, no study has addressed the temporal dynamics or the impact of climate factors on the risk of ZCL in the old world. The objective of this project is to analyze the temporal trends of *L.major* leishmaniasis in Sidi Bouzid using monthly data (1991-2007) and assess the impact of climatic factors on the risk of the disease. More specifically, this study aims to establish a model for prediction of ZCL and evaluate the relative impact of climatic factors on the risk of ZCL. The forecasting technique based on the method of Box and Jenkins showed that

historical observations and the random lagged by one month can almost perfectly predict the incidence of ZCL. Based on multivariate autoregressive models, the present study confirmed the relationship between cutaneous leishmaniasis of the old world and climatic factors.

**Keywords.** ARIMA, VAR, Multivariate AR, LCZ, Climate

**Introduction.**

La leishmaniose cutanée (LCZ) causée par Leishmania major constitue un véritable problème de santé publique dans les pays du Maghreb et l'Est de la région méditerranée[1]. Par conséquent, la prévision des épidémies et l'alerte précoce reste une priorité de la recherche pour les programmes de contrôle dans le monde ancien de leishmaniose cutanée en l'absence d'un vaccin efficace. Toutefois, pour cette menace de santé, ces systèmes ne peuvent pas être correctement conçus en l'absence d'évaluation fiable et valide de l'évolution temporelle des épidémies et de leurs déterminants.

**Matériel et méthodes.**

Afin d'atteindre cet objectif, des enregistrements mensuels du nombre de cas LCZ dans le gouvernorat de Sidi Bouzid (SBZ) dans le centre de la Tunisie, entre janvier 1991 et décembre 2007, ont été collectés auprès du programme national de contrôle de la leishmaniose (NCPL) de la Direction Régionale de la Santé de SBZ. Les données météorologiques mensuelles pour le gouvernorat de Sidi Bouzid entre 1991 et 2007 ont été fournies par l'Institut National de Météorologie. Elles englobent: température moyenne en degrés Celsius, humidité relative moyenne en pourcentage et précipitations cumulées en millimètre. L'exhaustivité de ces données météorologiques et des cas LCZ est parfaite. Les méthodes statistiques utilisées sont essentiellement des séries chronologiques. En premier lieu, on présente le modèle ou le processus ARIMA (Autoregrssive Integrated Moving Average) qui a été proposé par Box et Jenkins (1970)[2]. Cette procédure se distingue des autres techniques statistiques par l'amélioration de la qualité des prévisions et elle possède les étapes suivantes: désaisonnalisation, stationnarisation, identification, estimation, tests de validation et prévision.

Les modèles ARMA sont représentatifs d'un processus généré par une combinaison des valeurs passées et des erreurs passées. Ils sont définis par l'équation: ARMA *(p,q)*:

$$y_t - \theta_1 y_{t-1} - \theta_2 y_{t-2} - ... - \theta_p y_{t-p} = \varepsilon_t - \alpha_1 \varepsilon_{t-1} - \alpha_2 \varepsilon_{t-2} - ... - \alpha_q \varepsilon_{t-q}$$

L'existence d'un facteur de saisonnalité dans une série est basée sur un ensemble de tests [3, 4, 5, 6]. En premier lieu, il y a le test de saisonnalité stable qui est un test d'analyse de la variance à un facteur. On dispose de $k$ échantillons (les estimations de la composante saisonnière-perturbation, $k = 12$ mois) de tailles respectives $n_1$, $n_2$,..., $n_k$ et on suppose que la saisonnalité influe uniquement sur les moyennes des distributions et non sur leur variance. Il s'agit donc d'un test d'égalité des $k$ moyennes $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$. Si on considère chaque échantillon comme issu d'une variable aléatoire $X_j$ suivant une loi de moyenne $m_j$ et d'écart-type $\sigma$, le problème est de tester:

$$H_0 : m_1 = m_2 = ... = m_k$$

$$H_1 : m_p \neq m_q \text{ pouraumoinsuncouple(p,q)}$$

L'équation dite d'analyse de la variance s'écrit:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2 = \sum_{j=1}^{k} n_j(\bar{x}_{.j} - \bar{x}_{..})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{.j})^2$$

La variance totale se décompose donc en variance des moyennes, variance due au facteur saisonnalité, et en une variance résiduelle, c'est à dire,

$$S^2 = S_S^2 + S_R^2$$

Si l'hypothèse $H_0$ est vraie, on montre que la quantité

$$F_S = \frac{S_S^2/(k-1)}{S_R^2/(n-k)}$$

suit une loi de Fisher $F\,(k-1;\, n-k)$ à *k-1* et *n-k* degrés de liberté. Si la quantité calculée est supérieure à la valeur critique d'une loi de Fisher, on conclut à une influence significative du facteur saisonnalité (les moyennes mensuelles ne sont pas toutes égales).

En second lieu, il y a le test de saisonnalité évolutive qui est basé sur un modèle d'analyse de la variance à deux facteurs (le mois) proposé par Higginson (1975). Ce test repose sur la modélisation des valeurs de la composante saisonnière-perturbation:

$$X_{ij} = b_i + m_j + \varepsilon_{ij}$$

où: $m_j$ désigne l'effet du mois *j (j=1,. . . ,k).*

$b_i$ désigne l'effet de l'année *i (i=1,. . . ,N)* où $N$ est le nombre d'années complètes.

$\varepsilon_{ij}$ représente l'effet résiduel, réalisation de lois indépendantes et identiquement distribuées de moyenne nulle. Le test est basé sur la décomposition suivante:

$$S^2 = S_{IM}^2 + S_{IA}^2 + S_R^2$$

$$S^2 = \sum_{j=1}^{k}\sum_{i=1}^{N}(X_{ij} - \bar{X}_{..})^2 \text{ avec } \bar{X}_{..} = \sum_{j=1}^{k}\sum_{i=1}^{N}\frac{X_{ij}}{(kN)}$$

$$S_{IM}^2 = N\sum_{j=1}^{k}(\bar{X}_{.j} - \bar{X}_{..})^2 \text{ avec } \bar{X}_{.j} = \sum_{i=1}^{N}\frac{X_{ij}}{N}$$

$$S_{IA}^2 = k\sum_{j=1}^{N}(\bar{X}_{i.} - \bar{X}_{..})^2 \text{ avec } \bar{X}_{i.} = \sum_{j=1}^{k}\frac{X_{ij}}{N}$$

$$S_R^2 = \sum_{j=1}^{k}\sum_{i=1}^{N}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} - \bar{X}_{..})^2$$

$S_{IM}^2$ est la somme des carrées « Inter Mois », $S_{IA}^2$ est la somme des carrées « Inter Années » et $S_R^2$ est la somme des carrées « Résiduelle ». Soit l'hypothèse $H_0^* : b_1 = b_2 = \ldots = b_N$, c'est-à-dire que la saisonnalité n'évolue pas au cours des années, peut être testée grâce à la statistique suivante:

$$F_M = \frac{S_{IA}^2/(N-1)}{S_R^2/(N-1)(k-1)}$$

qui suit, sous $H_0^*$, une loi de Fisher à $(N-1)$ et $(k-1)$ degrés de liberté.

Par la suite on passe au test non paramétrique de Kruskal-Wallis. Les données étant supposés dériver de $k$ échantillons indépendants $A_1$, $A_2, \ldots$, $A_k$ de tailles respectives $n_1$, $n_2, \ldots$, $n_k$. Le test est fondé sur la statistique suivante:

$$W = \frac{12}{n(n+1)}\sum_{j=1}^{k}\frac{S_j^2}{n_j} - 3(n+1)$$

Où $S_j$ est la somme des rangs des observations de l'échantillon $A_j$ dans la série des $n = \sum_{j=1}^{k} n_j$ observations. Cette quantité suit, sous l'hypothèse nulle, une loi du Chi-deux à $k-1$ degrés de liberté. Le test de présence d'une saisonnalité identifiable est construit à partir des valeurs des statistiques de Fisher des tests de saisonnalité stable (statistique $F_S$) et de statistique évolutive (statistique $F_M$) évoqués ci-dessus. Ce test a été élaboré, à partir de considérations théoriques et pratiques, par Lothian et Morry (1978) [5]. La valeur de la statistique de test $T$ s'exprime comme suit: $T = (\frac{T_1+T_2}{2})^{1/2}$ avec $T_1 = \frac{7}{F_s}$ et $T_2 = \frac{3F_M}{F_s}$

Si $T > 1$, l'hypothèse nulle est rejetée et on conclut qu'aucune saisonnalité identifiable est présente.

Une fois que le test est significatif, on utilise la méthode X-12-ARIMA pour effectuer la désaisonnalisation. Cette technique repose sur un principe itératif d'estimation des différentes composantes par des moyennes mobiles adéquates. Ces dernières sont des outils de lissage conçus pour éliminer une composante indésirable de la série.

Une moyenne mobile est une combinaison linéaire d'opérateurs retard $L$ tel que

$M = \sum_{-m_1}^{m_2} \theta_i L^{-i}$, où $m_1$ et $m_2 \in \mathbb{N}$.On peut écrire cette combinaison sous une autre forme de la manière suivante:

$$M = L^{m_1}\sum_{i=0}^{m_1+m_2}\theta_{i-m_1}L^{-i} = L^{m_1}\sum_{i=0}^{m_1+m_2}\theta_{i-m_1}F^i = L^{m_1}\Theta(F)$$

où $\Theta(.)$ est le polynôme caractéristique de $M$, de degré $m_1 + m_2$ et d'ordre $m_1 + m_2 + 1$ et $F = L^{-1}$.

La détection de la stationnarité et de sa nature a été étudiée via le test de racine unitaire (Dickey-Fuller augmenté), le test de Ljung-Box et le test Phillips-Perron. L'estimation des ordres $p$ et $q$ d'un éventuel modèle ARIMA, est basée sur les coefficients de corrélations canoniques appelé méthode SCAN (smallest canonical correlation). Cette méthode permet d'identifier les ordres d'un processus stationnaire ou non stationnaires ARMA. Tsay et Tiao (1985) [7] ont proposé la technique, et Box et al. (1994) et Choi (1992) [8, 9] ont proposé une description de l'algorithme. Pour les tests de validation, il y a les tests de significativités des coefficients, coefficients de détermination ($R^2$) et le test de bruit blanc. La prévision via un modèle ARMA se fait par l'application du théorème de Wold au processus $X_t$ et considérons la forme $MA(\infty)$ correspondante:

$$X_t = \sum_{j=0}^{\infty} \pi_j \varepsilon_{t-j} \text{ et } \pi_0 = 1$$

Il s'ensuit que la meilleure prévision que l'on peut faire de$X_{t+1}$ compte tenu de toute l'information disponible jusqu'a la date $t$, notée $\hat{X}_t(1)$, est donnée par:

$$\hat{X}_t(1) = E(X_{t+1}/X_t, X_{t-1}, X_{t-2}, ..., X_0) = E(X_{t+1}/\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_0) = \sum_{j=1}^{\infty} \pi_j \varepsilon_{t-j}$$

Plus généralement pour une prévision à un horizon $r$ on a:

$$\hat{X}_t(r) = \sum_{j=1}^{\infty} \pi_j \varepsilon_{t+r-j}$$

$$X_{t+r} - \hat{X}_t(r) = \sum_{j=0}^{r-1} \pi_j \varepsilon_{t+r-j}$$

On peut déterminer un intervalle de confiance sur la prévision $\hat{X}_t(r)$, sous l'hypothèse de normalité des résidus. On montre alors que:

$$\frac{X_{t+r} - \hat{X}_t(r)}{var[X_{t+r} - \hat{X}_t(r)]^{1/2}} \hookrightarrow N(0,1)$$

Or on sait que:

$$E\{[X_{t+r} - \hat{X}_t(r)]^2\} = E[(\sum_{j=0}^{r-1} \pi_j \varepsilon_{t+r-j})^2] = \sum_{j=0}^{r-1} \pi_j^2 \sigma_\varepsilon^2$$

D'où

$$\frac{X_{t+r} - \hat{X}_t(r)}{\sigma_\varepsilon [\sum_{j=0}^{r-1} \pi_j^2]} \hookrightarrow N(0,1)$$

Pour la prévision multivariée, on présente le modèle VAR et les modèles AR multivariée. Ces deux techniques ont été envisagées dans le but de comparer une technique largement répandue dans le domaine économique (VAR) avec les modèles AR multivariée qui semblent plus liée à la réalité épidémiologique [10, 11].

Par définition, un modèle VAR est un processus vectoriel $x_t, t \in \mathbb{Z}$, de dimension $(n, 1)$, admet une représentation VAR d'ordre $p$, notée VAR($p$) si:

$$x_t = c + \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + ... + \Phi_p x_{t-p} + \varepsilon_t$$

ou de façon équivalente: $(I_n - \Phi_1 L - \Phi_2 L^2 - ... - \Phi_p L^p) x_t = c + \varepsilon_t \iff \Phi(L) x_t = c + \varepsilon_t$

Où c de dimension $(n, 1)$ désigne un vecteur de constantes, les matrices $\Phi_i$ de dimension $(n, n)$ , satisfont $\Phi_0 = I_n$et $\Phi_p \neq 0_n$. Le vecteur $(n, 1)$ des innovations $\varepsilon_t$est i.i.d $(0_n, \Omega)$ où $\Omega$ est une matrice $(n, n)$ symétrique définie positive.

$$E(\varepsilon_t) = 0$$

$$E(\varepsilon_t \varepsilon_\tau^{'}) = \left\{ \begin{array}{ll} \Omega, & t = \tau \\ 0, & t \neq \tau \end{array} \right\}$$

L'identification de l'ordre du processus VAR a été étudiée via plusieurs méthodes. La procédure de sélection de l'ordre de la représentation consiste à estimer tous les modèles VAR pour un ordre allant de 0 à h fixé de façon arbitraire ou fixé. Pour chacun de ces modèles, on calcule les fonctions AIC, SBIC, HQIC, ainsi que le critère FPE (Final Prediction Error).

**Application.**

Le meilleur modèle retenu par le modèle ARMA était le modèle ARMA(1,1) (tableau 1)

Table 1: Estimation et degré de signification des coefficients du meilleur modèle ARIMA(1,0,1)

| | Estimation | Erreur standard | Statistique du Student | Pr(>\|t\|) |
|---|---|---|---|---|
| ar1 | 0,508 | 0,093 | 5,451 | <0,001 |
| ma1 | 0,197 | 0,105 | 1,871 | 0,063 |

La prévision dynamique consiste à effectuer la prévision d'une valeur au temps$t$ en utilisant toutes les observations précédentes. Autrement dit, pour chaque temps $t \in 150$ (juin 2003),. . . , 204 (décembre 2007) on estime les coefficients du modèle en utilisant les observations de la série de 1 à $t$–1 et on prédire la valeur au temps $t$ en utilisant l'écriture du modèle. On remarque que la prévision dynamique par ce modèle $ARMA(1, 1)$ montre une grande performance de prévision. En fait, aucune valeur de la série LCZ se retrouve en dehors de la bande de confiance. Ce taux d'erreur $(0/54 = 0\%)$ correspond au risque de première espèce que l'on s'est fixé (5%). La stabilité du modèle a été

confirmée par la densité de chaque composante ($AR(1)$ et $MA(1)$) telle qu'elle est estimée par la méthode des noyaux. Nous avons utilisé le noyau gaussien et un choix de la fenêtre par validation croisée. Le modèle VAR retenu était celui $VAR(1)$. La seule variable significative était la variable LCZ avec un retard. Toutes les autres variables étaient non significatives. Ce modèle confirme l'importance du facteur AR déjà trouvé dans le modèle ARMA. Ce résultat mis en cause de considérer que les variables climatiques et LCZ sont endogènes. En revanche, l'utilisation du modèle AR multivariée, a mis en évidence et a confirmé l'importance des variables climatiques [12]. En se basant sur le critère d'AIC et les fonctions d'autocorrélations partielles, le modèle retenu contient les variables retardées suivantes: LCZ(1), LCZ(13), LCZ(14), humidité (0), humidité (1), pluviométrie (11) et pluviométrie (12). Ce modèle présente une interprétation épidémiologique. En fait, ces résultats confirment l'effet annuel de pluviométrie ainsi que l'effet significatif de l'humidité pendant la période de transmission.

**Conclusion.**

L'objectif du présent projet est d'analyser les tendances temporelles de la leishmaniose à *L.major* à Sidi Bouzid en utilisant des données mensuelles (1991 - 2007) et d'évaluer l'impact des variables climatiques sur le risque de la maladie. Plus spécifiquement, cette étude vise à établir un modèle de prévision de la LCZ et d'évaluer l'impact relatif des facteurs climatiques sur le risque de la LCZ. Dans le but d'atteindre nos objectifs, plusieurs techniques statistiques ont été utilisées. Dans un premier temps, nous avons utilisé l'algorithme de désaisonnalisation X-12-ARIMA qui a permis d'élucider un facteur saisonnier significatif de la série LCZ et une période inter-épidémique variant de 4 à 7 ans. La technique de prévision basée sur la méthode de Box et Jenkins a montré que l'historique des observations ainsi l'aléa décalés par un seul mois peuvent prévoir quasi- parfaitement l'incidence de la LCZ incluant sa distribution, sa tendance ou sa saisonnalité. La présente étude a confirmé la relation entre la leishmaniose cutanée de l'ancien monde et les facteurs climatiques. Les résultats, malgré leur richesse et innovation, ont clairement montré que des recherches supplémentaires sont nécessaires dans les zones endémiques pour développer des systèmes de suivi qui permettraient de prédire l'impact des changements climatiques sur le risque de la leishmaniose. Cette étude constitue un pas dans le développement des modèles d'alerte précoce de l'émergence des épidémies de LCZ.

# References


[1] Ozbel Y, Turgay N, Ozensoy S, Ozbilgin A, Alkan MZ, et al. (1995) Epidemiology, diagnosis and control of leishmaniasis in the Mediterranean region. Ann Trop Med Parasitol 89 Suppl 1: 89-93.

[2] Box G, Jenkins G (1970) Time Series Analysis, Forcasting and Control. Bibliographie 73 Holden-Day.

[3] Shiskin J, Allan HY, John CM (1967) The X-11 Variant of the Census-Method II Seasonal Adjustment. Document technique no15 du US Department of Commerce du Bureau of the Census.

[4] Findley DF, Brian CM, William RB, Mark CO, Bor-Chung C (1998) New Capabilities of the X-12-ARIMA Seasonal Adjustment Program (with Discussion). Journal of Business and Economic Statistics 16: 127-177.

[5] Lothian J, Morry M (1978) A Test for the Presence of Identifiable Seasonality when using the X-11 Program. Statistique Canada Document de recherche n_ 78-10-002E, Seasonal Adjustment and Time Series Staff

[6] Dagum EB (1988) The X11ARIMA/88 Seasonal Adjustment Method. Foundations and User's Manual, Time Series Research and Analysis Division Rapport technique de Statistique Canada.

[7] Tsay R, Tiao G (1985) Use of Canonical Analysis in Time Series Model Identification. Biometrika 72: 299-316.

[8] Box GEP, Jenkins GM (1994) Time Series Analysis: Forecasting and Control. Third Edition, Englewood Cliffs, NJ: Prentice Hall: 197-199.

[9] Choi B (1992) ARMA Model Identification. New York: Springer-Verlag: 129-132.

[10] Hamilton, J. (1994), Time Series Analysis, Princeton University Press, Princeton.

[11] Lütkepohl, H. (2006), New Introduction to Multiple Time Series Analysis, Springer, New York.

[12] Toumi, A., Chlif, S., Bettaieb, J., Ben Alaya, N., Boukthir, A., Ahmadi, Z.E., Ben Salah, A., 2012. Temporal dynamics and impact of climate factors on the incidence of zoonotic cutaneous leishmaniasis in central Tunisia. PLoS neglected tropical diseases 6(5), e1633.