***Algerian Journal of Environmental Science and Technology***
***March edition. Vol.9. Nº1. (2023)***
***ISSN : 2437-1114***
***www.aljest.org***

# A Combined DE Algorithm with SARIMA for Modeling and Predicting the Incidence of Zoonotic Cutaneous Leishmaniasis in Msila Province, Algeria

N. Frissou[1]*, M.T. Kimour[2], S. Selmane[1]

[1]Laboratory of Fundamental Computer Science Operational Research Combinatory and Econometrics (L'IFORCE) Faculty of Mathematics
University of Sciences and Technology Houari Boumedienne Algiers, 16111 Algiers- Algeria.

[2]Environmental Research Center Annaba, 23005 Annaba-Algeria

*Corresponding author: n.frissou@yahoo.fr; nfrissou@usthb.dz; Tel.: +213 663 08 02 06.

## ARTICLE INFO

***Article History:***

*Received* **: 13/01/2021**
*Accepted* **: 03/06/2021**

***Key Words:***

*Zoonotic Cutaneous Leishmaniasis, Differential Evolution (DE ) Algorithm, Time Series, ARIMA, SARIMA*

## ABSTRACT/RESUME

***Abstract:*** *Time series forecasting is a valuable tool to recognize and control the behavior of various practical systems, based on the data in a certain period of time. One of the most widely used method in time series forecasting is ARIMA (AutoRegressive Integrated Moving Average), and SARIMA, which extends ARIMA to handle the seasonal data. However, ARIMA-SARIMA has a weakness in determining the optimal model. In this research, we present a combined approach of the differential evolution (DE) algorithm and SARIMA model (p, d, q)×(P, D, Q), allowing to calculate the smallest Akaike Information Criterion (AIC) value, which represents a quality metrics of the statistical model. In doing such combination, we show that better accuracy and more convergence speed up of the calculated ARIMA-SARIMA model can be obtained. The data used in the study are monthly data of the Zoonotic Cutaneous Leishmaniasis, from January 2013 to December 2020 in Msila Province, Algeria.*

## I. Introduction

As a growing field of interest, time series forecasting is playing an important role in many industrial ad research fields such as biomedical, economics, planning, meteorology and agriculture. One of the most widely used method in time series forecasting is Autoregressive Integrated Moving Average (ARIMA) [1-2]. It is composed of an autoregressive (AR) model and a moving average (MA) model. The ARIMA model is a powerful solution for forecasting when the data contains no seasonality information. When data exhibits seasonal behavior, the seasonal autoregressive integrated moving average (SARIMA) model is more suitable for forecasting than ARIMA. Nevertheless, although SARIMA gives satisfactory results in widely forecasting domains, the method to determine its parameters has a weakness in determining the optimal model [2-5]. Indeed, searching an optimal SARIMA model is a complex combinatorial optimization problem and has always been a bottleneck in the time series forecasting field. Therefore, a supporting algorithm is needed to optimize the SARIMA model.

This research combines two methods which are SARIMA [2-4] and an Improved Differential Evolution (IDE) Algorithm, which is our novel improved version of differential evolution algorithm using the concept of adaptation-based learning. In fact, the differential evolution algorithm [7] is among the most powerful metaheuristics algorithms, widely used for optimization problems, but it is still suffer from falling into local optimal solution. Therefore, concepts of adaptation-based learning have been appropriately introduced into the original DE to prevent it from falling into local optima, while inducing faster convergence and better accuracy. We have demonstrated such strengths using benchmarks functions and a case study of zoonotic cutaneous leishmaniasis (ZCL) diseases forecasting in Msila province, Algeria.

The IDE-SARIMA process searches for the smallest AIC value [4], which is the criteria to find the best SARIMA (p,d,q) (P,D,Q) models. The obtained optimized model is then used for forecasting. Through its model quality and accuracy, estimated by the Root Square Mean Error (RMSE), forecasting results are calculated. The results will be compared with the results of the Box-Jenkins process [1]. The IDE -SARIMA model is expected to improve the accuracy and convergence speed of forecasting.

The main structure of the paper is given in the following. Section 2 presents the original differential evolution optimization algorithm and the novel improvement we introduced to it. Section 3, describes the SARIMA process, and develops its combination with the improved DE. Section 4 offers results and discussion o applying the IDE -SARIMA model of ZCL time series data, from January 2013 to December 2020. Finally, section 5 gives conclusion and future work.

## II. Improving the differential evolution algorithm

Differential Evolution algorithm (DE) is one of the most widely used evolutionary algorithms for solving real-valued numerical optimization problems.

The DE optimization algorithm has been successfully applied to many practical fields due to its good achieved results and complete attack strategy [8-11]. Nevertheless, there are still some weaknesses that make it subject to local optimum problems. To tackle such weaknesses, we used a nonlinear control parameter strategy to improve convergence speed, increase population diversity, and obtain better accuracy.

### II.1. The original DE

In the original basic differential evolution (DE) algorithm, the first step consists of generating a initial random population. After that, for a number maxIter of iterations, an evolving process is performed on every individual (target) by DE operators (mutation and crossover ) to generate a trial vector, and the next generation population is built through a competition between target individual and offspring individual [7-11].

**Mutation**: At iteration g, for each target vector $X_i^g$, a mutant vector is generated according to the following:

$$V_i^g = X_{r1}^g + F.(X_{r2}^g - X_{r3}^g) \text{ , r1} \neq \text{r2} \neq \text{r3} \neq \text{i.} \quad (1)$$

Where g indicates the generation, the indexes r1 r2, and r3 are indices randomly chosen from the set {1, 2, …, N }. $V_i^g$ is the mutation vector of the ith target vector $X_i^g$. The range of the scaling factor F is in [0, 2].

**Crossover:** A trial vector is produced by applying a crossover operation on the mutation vector $V_i^g$ and the ith target vector $X_i^g$,. The used crossover operator is as follows:

$$U_{ij}^g = \begin{Bmatrix} V_{ij}^g \ \ if \ rand_j \leq CR \ or \ j = jrand \\ \\ X_{ij}^g \quad otherwise \\ \end{Bmatrix} \quad (2)$$

where (i ∈ [1, N] and j ∈ [1,D]). **rand$_j$,** is a uniformly distributed random number in [0,1]. The crossover rate **CR** ∈ [0, 1] controls the diversity of the population and is closely connected with exploration power. **jrand** is a uniformly distributed random integer ∈ [1, D]. It guarantees that at least one component of trial vector is inherited from the mutant vector.

**Selection:** After crossover, the fitness of the trial vector can be calculated using the optimization problem. The generated trial vector is evaluated and compared with the target vector as follows:

$$U_{ij}^g = \begin{Bmatrix} U_i^g & if \ f\left(U_i^g\right) \leq f\left(X_i^g\right) \\ X_i^g & otherwise \end{Bmatrix} \quad (3)$$

### II.2. The Improved DE (IDE) Algorithm

To further improve the exploitation and exploration capability of DE algorithm, its convergence rate, and to prevent its prematurity, a Novel Enhanced DE (IDE) algorithm is proposed in the present research. In IDE, Novel mutation scheme and an Opposite-Based Learning (OBL) [12] operator are proposed. The crossover and selection operators are the same as in the basic DE ("DE/rand/1").

Novel mutation scheme. In the standard DE algorithm, the mutation operator F is constant, which cannot take into account the global search ability and convergence speed. To tackle this weakness, we propose to modify the value of mutation operator F in a way to be larger in the early stage, and then gradually decreases. In doing so, we obtain better global search ability and enhance the convergence speed. Based on the above principles, a new adaptive mutation operator is proposed in IDE algorithm (Figure 1). The scaling factor F is calculated as follows:

$$\beta = ((1\text{-g})/\text{G})^{**}2 \quad (5)$$

$$\alpha = 1\text{-rand1}^{**}\beta \quad (6)$$

$$F = F_0 \times 2^{\alpha} \quad (7)$$

In these two formulas, F0 denotes the initial value of the mutation operator, g denotes the current iteration number, and G denotes the maximum iteration number. From formulas (5-7), we see that the variation operator F has a linear decreasing trend. Using the above formula of the scaling factor F, the new mutation strategy is as follows:

$$V_{ij}^{g} = X_{Best,j}^{g} + F*(X_{r2,j}^{g} - X_{r3,j}^{g}),\ \text{if rand2} < \delta \qquad (8)$$

$$V_{ij}^{g} = X_{r1,j}^{g} + F*(X_{Best,j}^{g} - X_{r3,j}^{g}),\ \text{if rand2} \geq \delta \qquad (9)$$

Where r1 ≠ r2 ≠ r3 ≠ i.

```
(1) initial IDE parameters: NP, Gmax, CR, and F0.
(2) Initialize the original population pop
(3) calculate the individuals' fitness values, sort the individuals based o the fitness, and determine the best individual (bestI)with its fitness (bestF)
(4) while ((G≤ Gmax) do
(5)  for each individual Xi in pop do
(6)     randomly select three individual I1, I2, I3
(7)     Calculate Fi with equations (5-7), and δ with the equation (10);
(8)     for j = 1,D
(9)        calculate Vij according to equations (8-9)
(10)       calculate $U_{ij}^{g} = \begin{cases} V_{ij}^{g} & \text{if } rand_j \leq CR \text{ or } j = jrand \\ X_{ij}^{g} & \text{otherwise} \end{cases}$
(11)       calculate $U_{ij}^{g} = \begin{cases} U_i^g & \text{if } f(U_i^g) \leq f(X_i^g) \\ X_i^g & \text{otherwise} \end{cases}$
(12)       replace Xi with Ui
(13)    end for  #D
(14)  end for     #N
(15) end while
```

***Figure 1.*** *Algorithm of the IDE.*

rand1 and rand2 are two random values between 0 and 1. $X_{Best}^{g}$ is the best solution in generation g. j is the jth element in the individual vector. δ is a selection probability defined at every generation as follows:

$$\delta = \left(\frac{g}{G}\right)^{\frac{1}{4}} \qquad (10)$$

## III. Time series modeling using IDE-SARIMA

### III.1. SARIMA

When the underlying seasonal nature of the time series data has to be considered, we use SARIMA model [4] to fit seasonal time series. ARMA (p,q) process has the following form :

$$y_t = c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q} \qquad (12)$$

$$z_t \sim BB(0, \sigma_z^2)$$

where c, a constant and $\phi_p \neq 0$ , $\theta_q \neq 0$

with the backshift operator,

$$y_t = c + \phi_1 B y_t + \cdots + \phi_p B^p y_t + z_t + \theta_1 B z_t + \cdots + \theta_q B^q z_t \qquad (13)$$

$$\Phi\ (B) y_t = c + \Theta(B) z_t \qquad (14)$$

A process $y_t$ is a SARIMA(p,d,q)x(P,D,Q) if it verifies :

$$\Phi_s(B^s)\Phi(B)(1-B)^d(1-B^s)^D y_t = c_1 + \Theta(B)\Theta_s(B^s) z_t$$

Terms $(1-B)^d$ and $(1-B^s)^D$ expresses respectively the simple differentiation and the seasonal differentiation. A series is said to be d-integrated order if it has to be differentiated d times to obtain a stationary series. Likewise, a series is said to be seasonally integrated of order D if it has to be seasonally differentiated D times to obtain a stationary series. The order is therefore the order of simple differentiation, and D, the order of seasonal differentiation.

The differentiated series $(1-B)^d y_t$ is called the series of increases of y_t.

In fitting a SARIMA model, we perform data preprocessing, achieving stationarity, model identification, parameter estimation, diagnostic checking, and forecasting stages [2-6].

### III.2.Building the optimal model

In the analysis of the time series data, the precision of the model depends on the precision of its calculated parameters, to a great extent. It is worth noting that SARIMA has certain weakness in practical application, mainly manifested by algorithm prematurity and low convergence rate [4-6].

Firstly, the quality of a fitted model can be verified from the likelihood function value. The greater the likelihood function value, the better the fitting effect; as a general rule. Secondly, the number of unknown parameters in the model impacts such quality. Naturally, the smaller the number of unknown parameters the better, without significant precision change.

The minimum information test includes AIC and BIC. According to AIC criterion, the model which makes weighting function of fitting precision and parameter number reach the minimum value is considered relatively optimal, and AIC function has the form below:

$$\mathrm{AIC} = -2 * 1\mathrm{n}\ (\mathrm{L}) + 2\mathrm{n} \qquad (12)$$

Where, L stands for the maximum likelihood function value of the model, and n is the number of unknown parameters. SARIMA is still suffer from the difficult local optima problem, without an appropriate method to solve it. To tackle such challenge, we take the advantage of the superior characteristics of IDE, as efficient optimization scheme, to search optimal order of the SARIMA model, with respect of a give time series date.

In the sequel, the proposed IDE -SARIMA model is used to solve the problem of local optima, where the IDE is used at the initial stage to identify the parameters for SARIMA. Each individual of the population is composed of (p; d; q), and (P; D; Q) to represent SARIMA (p; d; q) ×(P; D; Q)s model, thus, an individual is encoded by real values. The process of determining optimal SARIMA model using IDE algorithm is as follows:

1. Selecting and preprocessing of the time series data.

2. Setting up the initial parameters of IDE: (N: population size, Lb, Ub, F0, CR , and maxIter)

3. Build the initial population (p, d, q, P, D, Q).

4. For each individual, calculate the fitness value, i.e. the AIC value according the equation (12). The smaller the value of the AIC, the higher the fitness.

5. Sort the individuals based on the fitness and determine the best individual bestI and its fitness bestfit.

6. Apply the evolving process from step (4) to (15) as depicted in the IDE pseudo-code.

7. Take the best individual bestI as the final optimal solution: (p*,d*,q*)x (P*,D*,Q*) for SARIMA model of the studied time series data.

Having the optimal SARIMA order (p*,d*,q*)x (P*,D*,Q*,s), we reuse IDE to determine optimal values of the coefficients using RMSE as fitness function.

## IV. A CASE STUDY: Zoonotic Leishmaniasis in Msila Province, Algeria

The Zoonotic Cutaneous Leishmaniasis (ZCL) Disease is a parasitic disease causing very debilitating skin or visceral conditions. It is a fatal disease if left untreated. It comes from various parasites of the Leishmania genus, transmitted by the bite of insects commonly called sandflies [15-17].

ZCL is distinguished by its rapid spread from old foci and is experiencing an increase in its incidence. This upsurge and the discovery of new foci make leishmaniasis a public health problem in Algeria. In this study, we are interested in the ZCL disease in the province of Msila, Algeria.

There were 96 notified ZCL incidences over the study period, from January 2013 to December 2020 in Msila province. Figure 1 shows monthly trend of ZCL incidence rate, indicating a seasonal pattern of the ZCL data through the period from January 2013 to December 2020. The peak of ZCL mainly occurred from October to February during the same epidemiological year (Figure 2) in the two years 2016 and 2017. Also, large values on the incidence of the disease were seen from November 2016 to February 2017 and from November 2017 to February 2018, but low values were recorded in the year 2013.

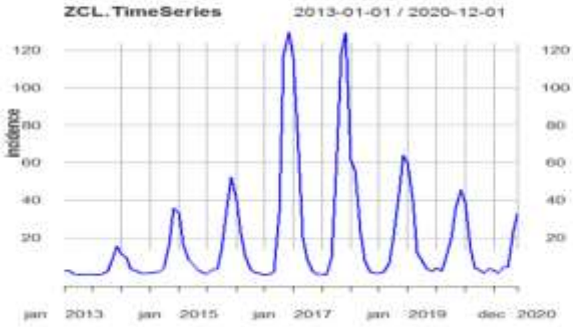

*Figure 2. Time series of the monthly reported ZCL incidences from 2013 to 2020.*

### IV.1. Modeling and Predicting the ZCL incidences using IDE-SARIMA

In order to identify the SARIMA model for the ZCL of Msila province-Algeria, the steps described by Box and Jenkins have been followed. For this purpose the data are partitioned into two stages. The training data is formed from the sample of observations starting from December 2020 to March 2021 and the testing data is concerned with the validity of model.

By applying the IDE-SARIMA on the ZCL time series data, we have obtained the best suitable SARIMA model: (1, 1, 1)×(2, 1, 0)12. This model can be used to make forecast time series (Figure 3). Due to the fundamental importance of forecast accuracy, a test should be performed to verify the forecasting accuracy, by comparing the forecast values with observational values. This test can also avoid under-fitting or over-fitting. Table 1 depicts the best suitable SARIMA model to ZCL time series data from January 2013 to December 2020, in Msila province, Algeria.

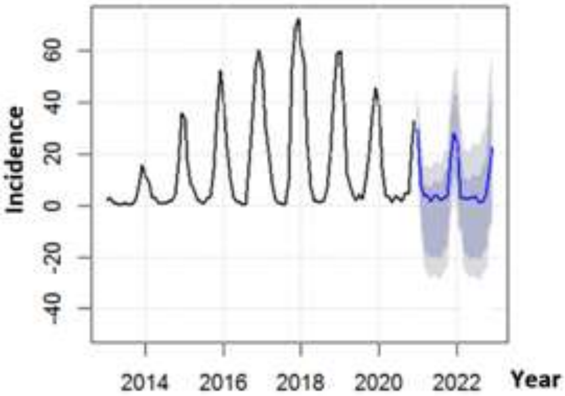

***Figure 3.*** *Prediction via the optimal SARIMA model, focusing on the dependent variable starting from December 2020 to December 2021.*

***Table 1.*** *The resulting SARIMA model.*

```
ARIMA(1,1,1)(2,1,0)[12]
Box Cox transformation: lambda= 0

Coefficients:
ar1      ma1      sar1     sar2
0.3029  -0.8964  -0.0846  -0.3282
s.e.  0.1421   0.0600   0.1288   0.1213

sigma^2 estimated as 0.3977:    log likelihood=-67.94
AIC=145.88 AICc=146.8 BIC=157.19,    RMSE=4.40
```

Data from January 2013 to December 2018 are used as the training set, while data from January 2018 to December 2020 are used as the testing set. The testselected model shows good forecasting accuracy of the RMSE equal to 4.40, which is relatively low and better than the obtained result by using the *auto.arima* function in the package “forecast” of statistical R language [18], which exhibited an RMSE equal to 6.20. Moreover, it is worth noting that our model presents consistency with the fact that more population of sandflies would result in a higher biting rate. Therefore, in the presence of reservoirs incorporating the parasites, this induces more infection transmissions. Sandflies in Msila are active from March until the end of September.

## V. Conclusion and future works

The research finds that the application of IDE algorithm in SARIMA Optimization model provides better result than the ARIMA Box-Jenkins model. The optimization of SARIMA Model with IDE algorithm on ZCL data has produced the best suitable SARIMA (1, 1, 1)× (2, 1, 0)12 model with AIC equal to 145.88 and RMSE equal to 4. 40. The time needed to search for SARIMA optimization model with IDE algorithm based on parameters tested is 0.21 minute for the provided ZCL data. The experimental results show that the optimization of SARIMA model using IDE algorithm is more optimal for finding the best model and more efficient in time. The IDE algorithm adjusts to minimize the number of generations and population in the search for the smallest AIC value. Thus, making it easier for analyst expert or user in determining the SARIMA model (p*,d*,q*)x(P*,D*,Q*) for forecasting and provides better assistance for faster decisions and policies making. Further researches on the optimization of SARIMA model with IDE algorithm could be elaborated based on the conceptual reflections drawn from this study.

## VI. References

1. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. Time series analysis: Forecasting and control 5th Edition *(*2015) ISBN: 978-1-118-67502-1.
2. Pickup, M. Fundamental concepts in time series analysis: Introduction to Time Series Analysis. SAGE Publications Inc (2015) 19–50.
3. Farsi, M.; Hosahalli, D.; Manjunatha, B.R.; Gad, I.; Atlam, E.; Althobaiti, A.; Elmarhomy, G.; Elmarhoumy, M.; Ghoneim, O.A. Parallel genetic algorithms for optimizing the SARIMA model for better forecasting of the NCDC weather data. *Alexandria Engineering Journal* 60 (2021) 1299–1316.
4. Al-Douri, Y.; Hamodi, H.; Lundberg, J. Time series forecasting using a two-level multi-objective genetic algorithm: A case study of maintenance cost data for tunnel fans. *Algorithms* 11(8) (2018).
5. Xiang, G.; Deyong, G.; Yuxuan, X. GA-ARIMA, Model-Based Analysis of Arrival Time at Bus Stop. *ICTETS 2020 Earth and Environmental Science* 587 (2020) 012050.
6. Imai, C.; Armstrong, B.; Chalabi, Z.; Mangtani, P.; Hashizume, M. Time series regression model for infectious disease and weather. *Environmental Research* 142 (2015) 319–327.
7. Storn, R.; Price, K. Differential evolution a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization* 11(4) (1997) 341-359.
8. Pan, J.S.; Liu, N.; Chu, S.C. A hybrid differential evolution algorithm and its application in unmanned combat aerial vehicle path planning. *IEEE Access* 8 (2020) 17691–17712.
9. Yang, X.S. Nature-Inspired Algorithms and Applied Optimization. *Studies in Computational Intelligence* (2018).

10. Tian, Y.; Tinghui, L. A New Adaptive Differential Evolution Algorithms. *Journal of Physics: conference series* (2020) 1437 012022.
11. Longlong, L.; Qisheng, Y. A new improved differential evolution algorithm. *Jiangxi Science* 04 (2017) 485-489.
12. Mahdavi, S.; Rahnamayan, S.; Deb, K. Opposition based learning: A literature review. *Swarm and Evolionary Computation* 39 (2018) 1–23.
13. Alfaki, M.M.A.; Masih, S.B. Modeling and Forecasting by using Time Series ARIMA Models. *International Journal of Engineering Research & Technology* 4(3) (2015).
14. Fang, X.; Liu, W.; Ai, J.; He, M.; Wu, Y.; Shi, Y.; Shen, W.; Bao, C. Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province China. *BMC Infectious Diseases* 20 (2020) 222.
15. Hong, A.; Zampieri, R.A.; Shaw, J.J.; Floeter-Winter, L.M.; Silva-Laranjeira, M.F. One Health Approach to Leishmaniases: Understanding the Disease Dynamics through Diagnostic Tools. *Pathogens* 9(10) (2020) 809.
16. Reithinger, R.; Dujardin, J.C.; Louzir, H.; Pirmez, C.; Alexander, B.; Brooker, S. Cutaneous leishmaniasis. *The Lancet Infectious Diseases* 7 (2007) 581–596.
17. Li, H. Predicting the number of visceral Leishmaniasis incidences in Kashgar, Xinjiang, China using the ARIMA-EGARCH model, *Asian Pacific Journal of Tropical Medicine* 13(2) (2020) 81-90.
18. Hyndman, R.J.; Khandakar, Y. Automatic Time Series Forecasting: the forecast Package for R, *Journal of Statistical Software* 27 (2008) Issue 3.

**Please cite this Article as:**

Frissou N., Kimour M.T., Selmane S., A Combined DE Algorithm with SARIMA for Modeling and Predicting the Incidence of Zoonotic Cutaneous Leishmaniasis in Msila Province, Algeria, ***Algerian J. Env. Sc. Technology***, **9:1 (2023) 2991-2996**