UNDER PEER REVIEW

Original Research Article

Prediction of cholera incidence by using the Autoregressive Integrated Moving Average model

ABSTRACT

Aims: The forecast is a topical subject, which aid in decision making and its performance. The aim of this study is to predict the disease between 1995 and 2010.

Place and Duration of Study: The choice of the disease is of after its appearance in our survey in the region of Gharb. Ttime series were illustrated between 1988-1994. Regional cholera annul data reported from ministry of health of Morocco.

Methods: The application of Box and Jenkins by the analysis of the series of cholera cases includes examining graphic series by using EVIEWR software, the consideration of the autocorrelation and partial autocorrelation functions, define the model that suits, estimate it, diagnose, the residue analysis and compare it with other models to choose the best for use in the forecasting process. Except the stationary series, we used IBMSPSS V22 for the other steps.

Results: Throughout this work, it is assumed that the underlying structure of the series follows an autoregressive integrated moving average (ARIMA) process. It is presumed that observations of the disease follow an autoregressive moving average process of order AR (1) and therefore ARIMA (1, 1, 0). The comparison of models of time series is extended away by using the statistics fit of the model: MAPE, BIC and R-squared, in addition to the sig. of the parameters and the analysis of residues by Ljung-Box and Durbinwatson statistic. The validation of the series is estimated by the calculation of the Mean Absolute Percentage Error (MAPE) and the signification of the parameter with P =0,05.

Conclusion: that the Brown model is the model of choice for the prediction of cholera cases.

Keywords: Cholera forecasting; ARIMA; MAPE; Stationary processes; Process non-stationary

1. INTRODUCTION

The forecast is one of the most important subjects to study in epidemiology, especially in climatic studies in relation to the diseases. This importance comes from its stake in the subject area of causation and the modeling of the disease. It is the process of prediction of future information using historical information, to nearly observe the behavior of diseases in the past [1,2].This reduced to capitalize on the new knowledge about infectious diseases, as easily as the capability improvement for the forecasting models. A chronological series is a set of observations xt, each of the observations is registered at a specific time t [3]. Cholera is among the diseases that are under surveillance by the Moroccan Ministry of Health. It's a contagious disease of bacterial origin, which is transmitted by dirty hands and the polluted water. The pathogenic agent is *V.cholera*e belonging to the family of the *Vibrioneceae* [4].

In our investigation, there are 20 cases reported in the survey of 1020 people questioned in the period between 2000 and 2010 [5]. According to the report of the Moroccan Ministry of Health (1994), the basin of Sebou is the area most affected by the disease in the state. The basin of the Sebou totaled 50% of the cases since the introduction of cholera. Moreover, the situation of the water hygiene and sanitation sector has not experienced significant improvements since 1993. In rural and peri-urban areas, hygiene and sanitation are far from the required standards. The level of economical aspect is a factor, which promotes the maintenance of diseases to cause, including hygiene practices in water supply for drinking and sanitation. Knowing that the rate of the lethality as a percentage of the cholera in Morocco varies between 7.4 for 1994 and 3.4 for 1990 [6]. According to the ministry of health that began the investigation in 1988, the region has the maximum number of cases in 1989, which were 665 cases among 3579 cases at the national level. The two cities SidiKacem and SidiSlimane show the utmost number of cases in 1989 those 425 cases among 3579 cases at the internal rate. In the

UNDER PEER REVIEW

same year, the Kenitra follows the two cities with 240 cases, while the total area is 665 shells. On that, point no cases were recorded since 1996.

The benefit of prediction is to understand the future potential occurrence of an outbreak. Preventive steps could be implemented to minimize its impact. The greater bulk of studies uses the ARIMA model for economic matters, but many studies used this model for prediction of diseases as leptospirosis [7], dengue [8], influenza H5N1 [8], suicide mortality [9], malaria and hepatitis A [10], and for other infectious diseases [11,12,13,14]. The aim of this study is to predict the disease between 1995 and 2010 by the application of the ARIMA process and develop disease surveillance systems to produce high quality, long-term data necessary for the development and testing of the model.

2. MATERIAL AND METHODS

Box and Jenkins [15,16,17,18] have proposed a forecast technique for univariate time series which is based on the concept of the ARIMA process. This technique has three steps: model identification, model estimation and model validation. Given a time series X_t, the ARMA model is a tool for understanding and predicting, possibly, the future values of this series. The model is framed of two parts: autoregressive part (AR) and moving average part (MA).The model is generally noted ARMA (p,q), where p is the order of the AR part and q the order of MA part. An autoregressive model and moving average of orders (p,q)(abbreviated as ARMA(p,q)) is a temporal process(X_t, $t \in \mathbb{N}$) checking:

$$Y_t=\varphi_p Y_{t-p} +\dots +\varphi_1 Y_{t-1}+\alpha_t-\theta_1 Y_{t-1} - \dots -\theta_q Y_{t-q}$$

With φ_i and θi are constants, and the terms of errors α_t are independent of the process. An autoregressive model *AR (p)* is an ARMA (p,0) and a moving average model *MA(q)* is an ARMA(0,q). The quality of a forecast largely depends on the choice of the one or the other of these methods. It likewise depends on "the art" of the forecaster to integrate a set of knowledge and to identify in a crowd of information those that are the most indicated to make predictions of audit evidence. The selection of method depends on several characters that are the following [19]:

1) The reduction of the size of the forecast errors, which must take into account the time of choice, follows by the choice of model. Sometimes the size of these errors is measured by the estimation of one of these three measures (MAD-MSE-MAPE).Mean Absolute Percentage Error (**MAPE**): it is the relative error made by a forecasting model. The following equation shows how one calculates this criterion:

$$\textbf{MAPE}= \left(\frac{100}{n}\right) \sum_{j=1}^{k} \left|\frac{yt-pt}{yt}\right| \qquad [20,21]$$

$\mathbf{yt}$: Initial observation, $\mathbf{t}$: forecast of the initial observation and **n**: number of initial observations;
The principle of the normalized **Bayesian Information Criterion** (BIC) criteria is to penalize the log-likelihood of estimated parameters associated with the data, reflecting the model fit, either by the number of independent parameters of the model, or the size of the sample. The best model in the sense of the criterion BIC is the one for which the value of the criterion is the lowest.
This criterion is defined in the following way:

$$\textbf{BIC = -2*LogLik + K*log(n)}$$

Or
LogLik represents the log-likelihood of the parameters associated with the data;
K refers to the number of independent parameters in the model;
N is the number of individual component the sample;
The elevation of the linear fit of the regression equation between the dependent variable Y and the set of explanatory variable is determined by R-squared:

$$R^2= SCR / SCT$$

With:

$$0 < R^2 < 1$$

Or
SCR = $\sum(\hat{y}t - y)^2$ **SCT** = $\sum(Yt - y)^2$
$\hat{y}$t is the estimated values, yt is the actual values, y is the average of the actual values.

UNDER PEER REVIEW

2) The type of the forecast request: If the point of the forecast is requested by the study, whereas the use of a traditional model and simple is sufficient to reach the goal. In several studies, the determination of the interval of the forecast is important and for this, we used the model of Box & Jenkins as a modern forecasting model.
3) The number of temporal observations: if the number of temporal observations is small then it is not necessary to apply a modern method.
4) The cost of the projection method used: This cost depends on the accessibility of statistical software.
5) The ease of the statistical tests and the understanding of the model chosen.
6) Effectiveness of the model that appears in the results of the forecast.

Analysis of the series of cholera cases includes examining graphic series by using EVIEWR software, the consideration of the autocorrelation and partial autocorrelation functions, define the model that suits, estimate it, diagnose, the residue analysis and compare it with other models to choose the best for use in the forecasting process. Except the stationary series, we used IBMSPSS V22 for the other steps [23].

3. RESULTS AND DISCUSSION

3.1 Initial review of the graph of the series

The series of cases of cholera disease includes seven observations, representing the historical development from 1988 to 1994. The first step is to determine whether the time series is in a stationary state with a mean plot of range. If non-stationary, it must be transformed in a serial temporal stationary by applying an appropriate degree of differentiation (d) to the data set. Input for ARIMA series must be fixed, with a constant mean, variance and autocorrelation in time. Through the graph in figure 1, there is a general upward trend over the period studied, with the decrease of the infection after the year 1989, which means that the series is stationary in the average arithmetic in general.

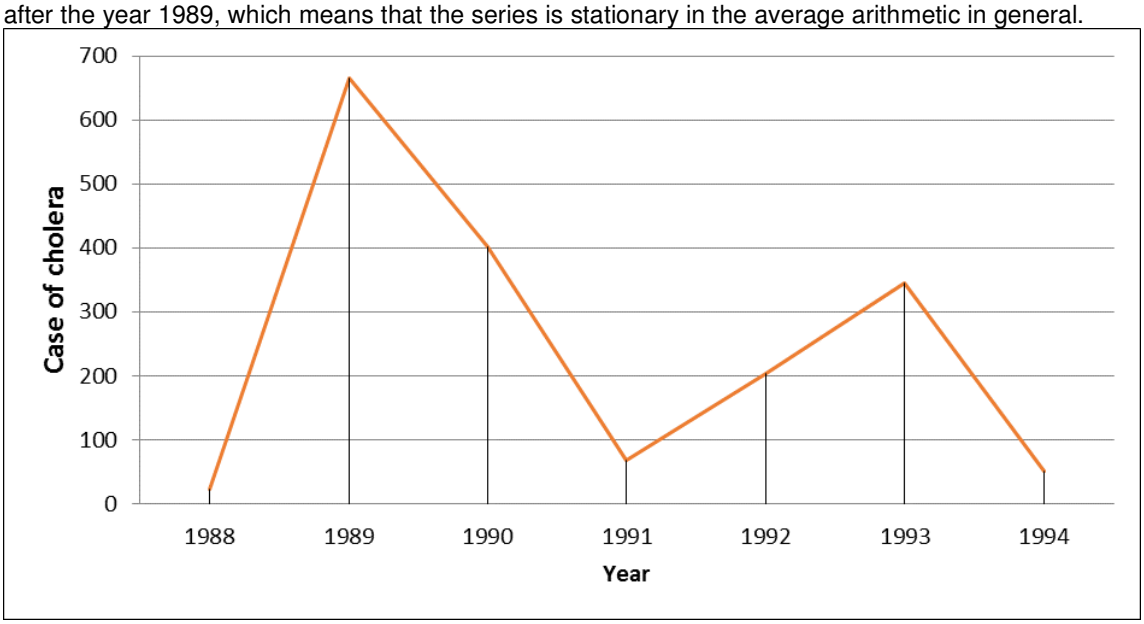
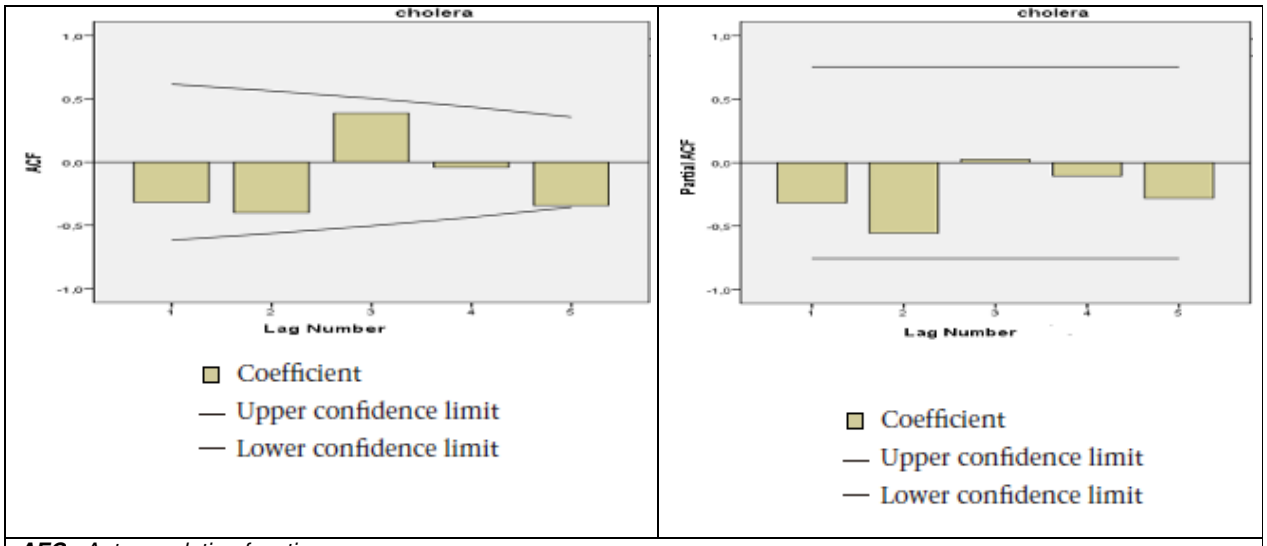


Fig. 1. Graphic of the series of the disease of cholera between (1988-1994)

The autocorrelation is a concept related to that of correlation: it is not only a calculation between two different chronological series but between the series and itself at different offsets in the time. Once this parameter is fixed, it is necessary to specify the order *p* of the auto-regressive process and *q* that of the moving average. The graphs of the function of autocorrelation (ACF) and the function of partial autocorrelation (PACF) allow depending on their aspects to correctly identify the parameters *p* and *q* whose values do not exceed two in general rule:
$P \in \{0,1,2\}$ et $q \in \{0,1,2\}$. The function of autocorrelation, recorded ACF, is constituted by the set of autocorrelations $P_k = corr\ (Y_t, Y_{t-1})$.
P of the series calculated for zero offsets order k, $k \in \{1, \ldots\ldots\ldots\ldots, K\}$

UNDER PEER REVIEW

The function of partial autocorrelation, recorded (PACF) is constituted by the set of partial autocorrelations, the coefficient of partial autocorrelation measuring the correlation between the variables between Y_t *and* Y_{t-k} , the influence of the variable being controlled for $i<k$. In addition to the correlation coefficients displayed the confidence intervals at 95 %, which allow determining what are the statistically significant coefficients to take into account [24]. We examine the autocorrelation and partial autocorrelation functions (fig.2). It was noted that the parameters of autocorrelation are in the confidence limit while they are significant.



AFC= *Autocorrelation function*

Fig. 2. Autocorrelation function and partial Autocorrelation function of cholera in Gharb area, Morocco

To verify the stationarity and formally confirm the previous findings, we estimate the augmented Dickey-Fuller test [25]. It was noted that the series is stationary by the EVIEWS software. We conclude that d = 1 in the ARIMA (p, d, q) model.

Table 1. Test of Dickey-Fuller by the EVIEWS software

Null Hypothesis: Unit root (individual unit root process)				
Series: SER01				
Date: 05/08/14 Time: 17:04				
Sample: 1988 1994				
Exogenous variables: Individual effects, individual linear trends				
Automatic selection of maximum lags				
Automatic selection of lags based on SIC: 0				
Total (balanced) observations: 7				
Cross-sections included: 1				

UNDER PEER REVIEW

Method			Statistic	Prob.**
ADF - Fisher Chi-square			4.20838	0.01219
ADF - Choi Z-stat			-1.16532	0.01219
** Probabilities for Fisher tests are computed using an asymptotic Chi				
-square distribution. All other tests assume asymptotic normality.				
Intermediate ADF test results UNTITLED				
Series	Prob.	Lag	Max Lag	Obs
CHOLERA	0.01219	0	0	7

3.2 Identification of the model

Once the parameter (d = 1) is fixed, it should specify the order of the Autoregressive process *p* and *q* of the moving average. The Correlograms (graphs of the function of autocorrelation and partial autocorrelation function) allow according to their aspects to correctly identify the parameters *p* and *q*. Through the (fig.2,3) of autocorrelation and partial autocorrelation functions, it seems that the latter stops clearly in the first and the second parameter of time and increases only in the middle and therefore the latter differs significantly from zero when α = 0.05. Although it seems that each parameter of partial autocorrelation after the second parameter of time differs significantly from zero at α = 0.05 level, but the autocorrelation function decreases gradually and was discontinued after a short time parameter. This may be enough proof that the random process follows the AR (1) model. It is a discreet temporal process (Xt, t ∈ ℕ) checking:

$$Y_t=\varphi_p Y_{t-p} + \dots + \varphi_1 Y_{t-1} + \alpha_t - \theta_1 Y_{t-1} - \dots - \theta_q Y_{t-q}$$

Where the parameters φi and θi are constants, and the terms of errors αt are independent of the process [26]. To be sure, we test the following statistical hypothesis:

$H_0 : \varphi_{11} = 0$ against $H_0 : \varphi_{11} \neq 0$

$SE(\varphi_{11}) = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{7}} = 0{,}377$

$Z = \frac{\varphi 11}{SE(011)} = \frac{0{,}240}{0{,}377} = 0{,}594 < 2$

It can be inferred that the correlation coefficient is significantly different from zero at the level α=0.05. After examination of the coefficients of partial autocorrelation, we find that φkk< 0.282 for each k = 2, 3, Thus, we can say that there is no reason or strong evidence that the function of partial autocorrelation interrupted the first time parameter, which supports the ability to use the model AR (1) and therefore ARIMA (1, 1, 0). The initial model of the series is written:

$Y_t = \varepsilon_t + \varphi Y_{t-1}$

Ɛt : white noise, φ: fixed amount is the main parameter of the model [27].

3.3 Model estimation

The ARIMA model is estimated to give a number of settings after choosing the values and use the non-linear method to estimate instead of the method of least squares. In this step, usually several converging models are estimated and compared where the parameters of the best model are significantly different from zero. It can also be compared to the total residue square as a measure of the quality of the model (Fig.3,4). In the table above we give the estimation of the parameters of the model ARIMA (1, 1,0) of the series of cholera. It is significantly different from zero array as φ_1 = 0.388 and constant= 0, 907. On notes that the setting of the ARIMA (1, 1, 0) model is not significantly as a consequence, we compare it with the larger models ARIMA (2, 1, 0) and (1, 1, 1), the

UNDER PEER REVIEW

lower model ARIMA (0,1, 0), model smoothing of Holt and Brown and the simple regression model (Table 3).The table 3 shows that models ARIMA (2, 1, 0), ARIMA (1, 1.1) and smoothing of Brown model convergent in words preferably with a preference in mean absolute percentage error. Thus, we compare them in the next step.

Table 2. Model parameters ARIMA (1, 1, 0) of the series of cholera

	Estimation	Standard error	t	Sig.
Constant	-14,71	118,396	-0,124	0,907
ARIMA	-0,446	0,461	-0,967	0,388

Table 1. Comparison of the ARIMA models with exponential smoothing and simple regression for the series of cases of cholera

Model	BIC	R^2	MAPE	Parameters Signification
ARIMA (1, 1, 0)	12,685	0.008	203,39	Constant =0. 907 φ_1=0.388
ARIMA (2,1, 0)	12,96	0,270	83,52	Constant=0.559, φ_1=0,361, φ_2=0.016
ARIMA (1, 1, 1)	13,10	0.162	127,07	Constant=0.530,φ_1=0.694, φ_2=0,988
ARIMA (0, 1, 0)	12,17	1,649	224,46	Constant=0,977
Model of Holt	11.713	0.911	252,433	alpha=0,717 , gamma=1
Model of Brown	11,632	0,870	192,950	alpha=0,176
Simple Regression	1,356	0.492	220,183	alpha=0,390

3.4 Model validation

All initial series contain 7 observations from 1988 to 1994. These observations served for the calibrations of the models (identification and estimation) while the assessment of the quality of the forecasts made in the last 7 following the forecast years (2003-2010). The validation was done by calculation of MAPE (Mean Absolute Percentage Error). We choose the Brown model.

Table 2. Models validation

		ARIMA (2, 1, 0)		ARIMA(1,1,1)		Brown Model	
Year	**Initial observation yt**	**Prediction of the initial observation pt**	**|(yt-pt)/yt|**	**Prediction of the initial observation pt**	**|(yt-pt)/yt|**	**Prediction of the initial observation pt**	**|(yt-pt)/yt|**
1988	23	6	0,739	168	6,304	176	6,652
1989	665	210	0,684	86	0,87	158	0,762
1990	401	66	0,835	67	0,832	147	0,633
1991	68	-114	2,676	29	0,573	136	1

UNDER PEER REVIEW

1992	204	31	0,848	-3	1,014	125	0,387
1993	344	36	0,895	-38	1,11	114	0,668
1994	51	-171	4,352	-7	1,137	103	1,019
			Sum=		**Sum=**		**Sum=**
			11,031		11,843		11,123
			MAPE=		**MAPE=**		**MAPE=**
			158,906		169,198		157,593

3.5 Analysis of the residuals of the model Brown

In the (Fig.3, 4), we note that the dispersal of residues (ACF and PACF) points are distributed in random shape around zero.

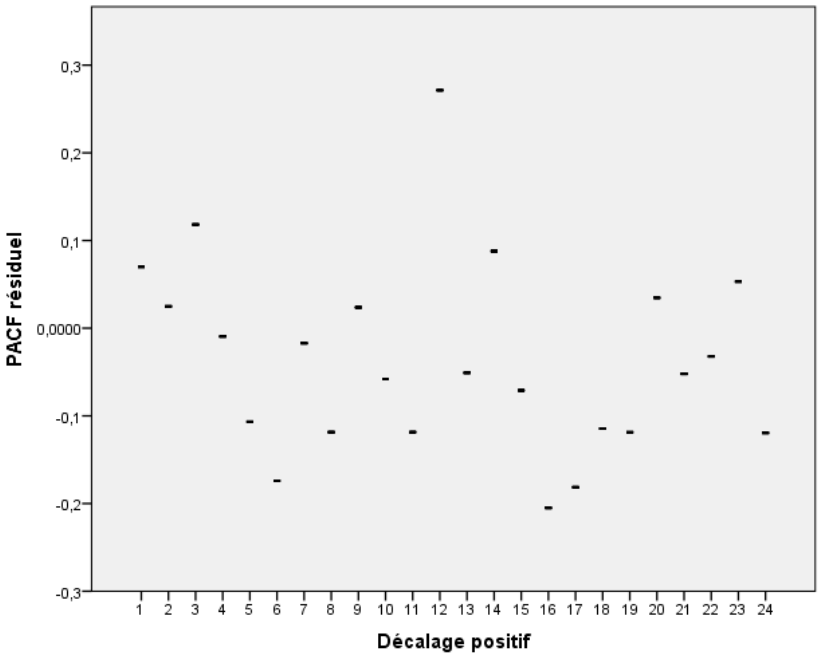


Fig. 3. Dispersion of points of residues PACF of the series of cholera

UNDER PEER REVIEW

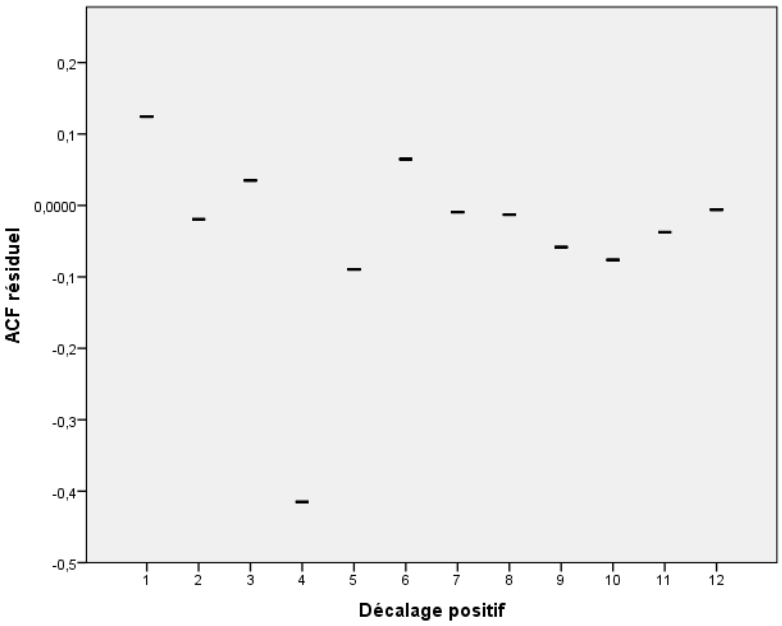


Fig. 4. Dispersion of points of residues ACF of the series of cholera

3.6 Forecasting

In the above table, we give the forecast of the series of values of cholera in the region of Gharb:

Table 5. Forecast of the series of cases of cholera in the Gharb

Year	Forecast
1995	170
1996	158
1997	147
1998	136
1999	125
2000	114
2001	103
2002	91
2003	80
2004	69
2005	58
2006	47
2007	36

UNDER PEER REVIEW

2008	24
2009	13
2010	2

4. CONCLUSION

The process of the study following the approach of modeling BOX-Jenkins was to detect the aberrations in onset of infectious diseases in history, so that effective interventions can be taken to prevent new cases and to control the disease. This helps to make decisions on the monitoring of disease and the management of risk as suggested by previous studies. The empirical results suggest that Brown model is the best and most accurate, which successfully identified the dynamics of cholera disease. Other studies have concluded the effectiveness of this model on medical research, [28], and the influenza mortality of the all [29].

The number of cases of the cholera disease is arriving at its maximum in 1989. By comparing the predicted values with the actual values, it is noted that the values provided by Brown model are directed toward a single trend and therefore to a decrease because of the lack of the number of observations such that 7 observations are not representative for the prediction by the ARIMA model (Fig.5). Benneyanet al.,2003 demonstrates how the control panels may be able to detect statistically significant signals in the patterns in the data more quickly than traditional statistical methods [30]. Since 1996, the number of disease incidences per year declined with a very predictable 11-12 disease occurrence per year (Fig.6).

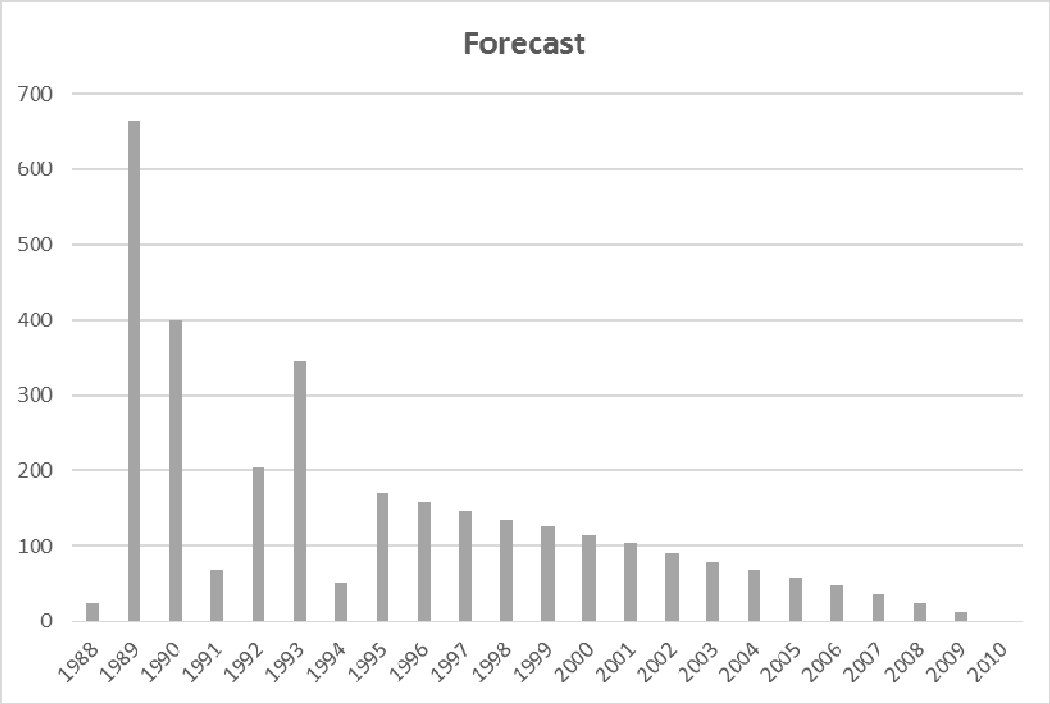
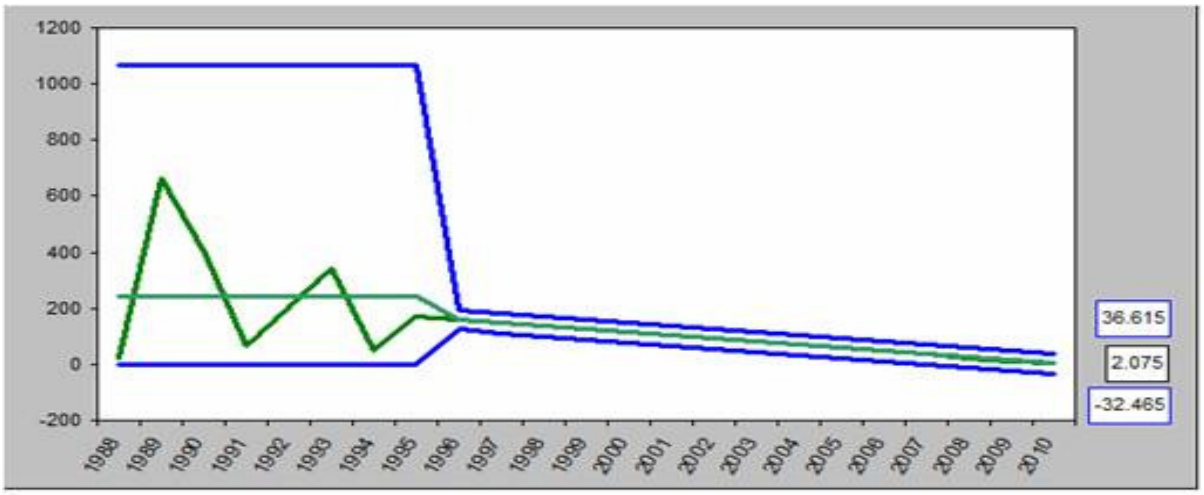


Fig. 5. Observed and forecasted Values of the cholera

UNDER PEER REVIEW



Decline from 1996 with a almost constant 11.16 diseases per year.

Fig. 6. Number of diseases changed over the years

REFERENCES

1. Dobson A., Kutz S., Pascual M., Winfree R. Pathogens and parasites in a changing climate. In: Lovejoy, L.H.a.T.E. (Ed.). Climate Change and Biodiversity: Synergistic Impacts. Yale University Press, New Haven. 2003:.33–38.
2. Gage, K.L., Burkot, T.R., Eisen, R.J., Hayes, E.B. Climate and vector-borne diseases. Am. J. Prev. Med. 2008:35, 436–450.
3. Peter H., Gleick. Dirty Water: Estimated Deaths from Water-Related Diseases 2000-2020. 2002: 1-12.
4. Berche P. Choléra et environnement. Journal de MCd Ma1 Infect. 1999 : (29):301-7.
5. Alwashali E., Jghalef B., Fadli M. Using the factorial correspondence analysis to determine the risks of water related diseases, hygiene and Good practices with survey data in Gharb area, Morocco. European Scientific Journal . 2014: 10 (3): 203-212.
6. Ministre de la Santé Marocaine. Bilan sur le choléra au Maroc. Documentation de la bibliothèque de la faculté des sciences à Rabat, Maroc. 1994. French.
7. Chadsuthi S., Modchang C., Lenbury Y., Iamsirithaworn S., Triampo W. Modeling seasonal leptospirosis transmission andits association with rainfall and temperature in Thailand using time-series and ARIMAX analyses.Asian Pac J Trop Med. 2012:5 (7):539–546.
8. Kan M.J., Price N., Scotch M., Rabinowitz P. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics. 2014:15:276.
9. Pridemore WA., Snowden AJ. Reduction in suicide mortality following a new national alcohol policy in Slovenia: an interrupted time-series analysis. AmJ Public Health. 2009: 99(5):915–920.
10. Nobre F.F., Monteiro A.B.S., Telles P.R., Williamson D., David G. Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology. Journal of Statist. Med. 2001:20:3051–3069.
11. Allard R. Use of time-series analysis in infectious disease surveillance. Bull World Health Organ. 1998:76(4): 327-333.
12. Helfenstein U. Box &Jenkins..Modelling of some viral infectious diseases. Stat Med. 1986: (1): 37-47.
13. Tong S., Hu W..Climate variation and incidence of Ross river virus in Cairns, ustralia: a time-series analysis. Environ Health Perspect. 2001:109(12): 1271-1273.

UNDER PEER REVIEW

14. Trottier H., Philippe P., Roy R. Stochastic modeling of empirical time series of childhood infectious diseases data before and after mass vaccination. Emerg Themes Epidemiol. 2006:3: 9.
15. Box G.E.P., Pierce D.A. Distribution of Residual Autocorrelations in Autoregressive Moving Average Time Series Models », Journal of the American Statistical Association. 1970 : (65): 1509-1526.
16. Box G.E.P., Jenkins G.M. Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco. 1976.
17. Box, G. E. P., Jenkins G. M., Reinsel G.C. Time series analysis: Forecasting and control, 3rd ed. Englewood Cliffs, N.J.: Prentice Hall. 1994.
18. Box G. E., Jenkins GM., Reinsel GC. Time series analysis: forecasting and control. New York: Wiley. 5th Edition. 2015 : 652.
19. Bourbonnais R. Prévision des ventes. Polycopié du produit multimédia. Université de Paris-Dauphine. 2001 :1-90. French.
20. Bloomfield P. Fourier analysis of Time Series: An Introduction. John Wiley, New York. 2000: (2):288.
21. Marie P. (s.d.). Prévision de la demande. Consortium de Recherche Université Laval. French.
22. Dom N., Hassan A., AbdLatif Z., Ismail R. Generating temporal model using climate variables for the prediction of dengue cases in Subang Jaya, Malaysia .Asian Pac J Trop Dis. 2012.:3(5): 352-361.
23. Aide IBM SPSS v22. Cases studies, Statistics Base, Forecasting option: Time Series Modeling and Forecasting. 2014. website : web:*http://192.168.1.15:52218/help/index.jsp?topic=%2Fcom.ibm.spss.statistics.tut%2Fspss%2Ftutorials%2Fintrotut2.html*
24. Ramsey F. Characterization of the Partial Autocorrelation Function. Journal Ann. Statist. 1974: 2(6):1296-1301.
25. Lagnoux A. A Two-Step branching splitting model under cost constraint for rare event analysis. Journal of Applied Probability. 2011:46 (2):429-452.
26. Majed B. Statistical forecasting methods-1. University of Roi Saoud, Arabie Saoudite. 2002.
27. Mostafa S. Introduction In modern analysis of time series- Scientific Publishing Center, King Abdulaziz University.2005.
28. Armitage P..Statistical methods in medical research. Oxford: Blackwell Scientific Publications. 1971:504.
29. Choi K., Thacker SB. An evaluation of influenza mortality Surveillance, 1962-1979. Time series forecasts of expected pneumonia and influenza deaths. Am J Epidemiol. 1981:113:215-26.
30. Benneyan JC., Lloyd RC., Plsek PE. Statistical process control as a tool for research and healthcare improvement. QualSaf Health Care. 2003:12:458-64.