SMGroup

SM Journal of Biometrics & Biostatistics

Research Article

Application of Different Time Series Models on Epidemiological Data - Comparison and Predictions for Malaria Prevalence

Ram Rup Sarkar[1,2]* and Chandrajit Chatterjee[3]

[1]*Chemical Engineering and Process Development, CSIR-National Chemical Laboratory, India*
[2]*Academy of Scientific and Innovative Research (AcSIR), CSIR-NCL Campus, India*
[3]*Ministry of Statistics and Programme Implementation, India*

Article Information

Received date: Nov 30, 2017
Accepted date: Dec 11, 2017
Published date: Dec 18, 2017

***Corresponding author**

Ram Rup Sarkar, Chemical Engineering and Process Development, CSIR-National Chemical Laboratory, India,
Tel: +91-20-2590 3040;
Fax: +91-20-2590 2621;
Email: rr.sarkar@ncl.res.in

Keywords Time series models; ARIMA; GARCH; Random walk; Malaria prevalence; Prediction

Abstract

Vector-borne diseases, such as Malaria, are major causes of human mortality in many areas of world, especially in the developing countries. Statistical and data-based models can provide an explicit framework to develop an understanding of infectious disease transmission dynamics. Application of different time series models to analyse and predict financial data as well as epidemiological data is of long interest to researchers. It is always interesting to see how the time series models that are extensively used in the analysis of financial data can be applied and extended to explain epidemiological data. In this paper, we have studied epidemiological data (malaria prevalence) related to Slide Positivity Rates and deaths due to *Plasmodium vivax*, using three major classes of time series models, namely Auto-Regressive Integrated Moving Average (ARIMA), Generalised Auto-Regressive Conditional Heteroskedastic (GARCH) and Random Walk. Our results show that as expected the chosen models fit excellently with the financial data but also show good potentiality to fit epidemiological data and provide excellent predictions. The results demonstrate the applicability of such time series models in epidemiology, specifically for malaria prevalence, where these models with appropriate choice of parameters have not been used extensively. As far as future prevalence pattern is concerned, the prediction of these models may help researchers and public health professionals to design control programmes for malaria.

Introduction

Application of mathematics and statistics to the study of infectious disease can be traced back to 1760 with the study on small pox - a fatal disease of that time [1] and subsequently different types of models have been developed to understand the disease transmission process and being developed till date [2-4]. In recent years, analyses of mathematical models and comparisons with incidence data have made it possible to explore the relative benefits of the theoretical studies on immunization strategies [2,5-8] and estimation of critical vaccination level to eradicate an infection [9]. Statistical or Data-based models that fit curves of past temporal prevalence of a disease, does not make any assumptions about the internal mechanisms that a mathematical model provides and hence, become more popular among researchers in infectious diseases. Several statistical or data based models have been used to model different infectious diseases, such as, SARS outbreak [10,11] Hepatitis A virus infection [12], ovine Johne's disease [13], HIV [14] etc. But it is in malaria, where most of the data-based and time series modeling approaches are observed apart from standard mathematical models [4].

Malaria caused by *Plasmodium* protozoan is the most important tropical parasitic disease of humans for centuries, remaining widespread throughout the world. Current estimates describe the annual global burden of malaria as: 300-500 million cases, 1.1 million deaths and 44 million cases of disability [15] and it is a public health problem in several countries, including sub-Saharan Africa, Indian sub-continent, south-east Asia and Oceania and the Americas [16,17]. Despite the introduction of control programs in many parts of the world over the past few decades, the impact of malaria on human populations continues to increase. Epidemiological research on malaria is largely based on two distinct measures of parasite abundance within communities of people. The first is the incidence of infection or disease and the second measure is the prevalence of infection or disease. The measurement of incidence or prevalence is often based on the stratification of the population under study with respect to a variety of factors such as age, sex, social factors, environmental variability etc as is found in the sources [18,19]. Regression analysis has been used extensively to understand how disease prevalence changes based on other variables such as, environment, etc. Suitable models have been fitted to see the incidence pattern of the disease using an extensive data set for studying the climatic suitability in malaria transmission in Africa through MARA project [20] and similarly in Europe by Kuhn et al [21]. Subsequently several other techniques were used

OPEN ACCESS

ISSN: 2573-5470

to study the disease pattern under the influence of environmental factors, such as, Logistic regression modeling [22]; Poisson regression modeling in Indonesia [23]; Binary logistic regression modeling with fractional polynomial transformations [19]. In recent studies, researchers have used techniques such as time series analysis to show seasonality pattern in the malaria incidence [24] and Monte-Carlo simulation methods to model risk factors [25-27]. However, there exist drawbacks, which affect the suitability of these models being fitted into the incidence pattern of the disease. A recent study by Chatterjee and Sarkar [28], attempted to deal with these drawbacks by developing a simple non-linear regression methodology in modeling and forecasting malaria prevalence in Chennai city, India. Though much of research has been dedicated to the understanding of this disease and its probable influences, yet a robust and unanimous approach has remained elusive.

Finance and Epidemiology are two distinct disciplines that have evolved in parallel and have very low similarities with respect to statistical and mathematical modeling. Hence, it is always interesting to see if the time series models that are used extensively in the analysis of financial data may be used to analyze epidemiological data so as to develop short term forecasting models which can help in obtaining good predictions. Application of time series models to analyse financial data has a long history [29-35]. Several models were developed for predicting financial asset returns [36-38] and a class of models namely Generalized AutoRegressive Conditional Heteroskedastic (GARCH) models were used extensively in studying the volatility clusters with appropriate modifications and extensions. Researchers in other fields have also utilized the properties of this family of models, e.g. speech signal modeling in time-frequency domain [39]. Similarly for modeling time series data, AutoRegressive Integrated Moving Average (ARIMA) class of models and Random Walk model have been used extensively in almost all fields [10,12,40,41].

The major use of these models is to fit past data and estimate the future. This capability of a model improves the credibility of the underlying hypothesis of the model. Thus in a non-deterministic scenario, if we need to forecast the future behavior of a system based on past observations, these models provide the best predictions, leading to wide acceptance of such models in analyzing epidemiological data. Epidemiological data show a seasonal pattern with a long term trend and seasonal fluctuations [2], while financial data is in general fat-tailed, characterized by higher kurtosis and clusters of volatility between groups of data [42]. The inherent differences in the two datasets as explained above lead to the fact that models used for analyzing financial data sets are very scarcely used in epidemiological studies, specifically for malaria prevalence data.

The main aim of this paper is to demonstrate that with a greater understanding of the data, the established models of predicting asset returns can be used for predicting and forecasting vector borne disease dynamics (malaria, in our case). To establish this, we have chosen three general classes of models, namely the ARIMA, the GARCH and the Random Walk models. The results of the paper will also demonstrate that while the aforementioned models are applicable to the study of epidemiological datasets, some other models specifically, Geometric Brownian motion - a diffusion process with continuous time domain may be inappropriate for studying the same. Suitability of the chosen class of time series model in analyzing financial data has been demonstrated on a test data set obtained from daily stock prices. On the basis of the best fitted model, predictions of prevalence of the disease in a comparable time horizon has also been attempted along with goodness of fit results.

The epidemiological data consists of two types of time series-one a shorter time series of 12 time points (monthly data for one year) and the other, a considerably longer time series of 36 time points (monthly data for three years). This is to demonstrate comprehensively whether the selected models work well for shorter as well as longer time series. We have also performed suitable analysis for testing the between-groups homoskedasticity for all the data sets and diagnostic tests to assess the goodness of fit for the time series models. Our study reveals that some time series models, which are historically proven to work well for modeling asset returns, can be used successfully to analyse the behavior of epidemiological time series and to predict the disease prevalence in both shorter and longer time scales.

Materials, Methods and Models

Overview of the data

The epidemiological data, (i.e. malaria prevalence) is collected from the Vector Control office, Malaria section of the Municipal Corporation of Chennai office of the Government of Tamil Nadu, India. The data consists of two types of time series: (i) a short time series with 12 time points consisting of *Plasmodium Vivax* (PV) deaths in Chennai, from January 2006 to December 2006 and (ii) a comparatively longer time series with 36 time points consisting of Slide Positivity Rates (SPR = No. of positive cases detected / No. of Blood smears collected) of malaria prevalence in Chennai, from January 2002 to December 2004. During the period of January 2002 to December 2004, the Vector Control Office in Chennai has covered a population ranging from 4.27 mn to 4.42 mn. The number of blood smears collected lies in the range of 16306-63717 depending on the seasons. The number of cases tested positive are in the range of 1184-4275, out of the blood smears collected. Further details of the data collection are available in Chatterjee and Sarkar [28]. More recent data could not be obtained for the purpose of the study owing to archaic methods of data keeping in the Municipality and lack of updated data.

Preliminary Testing of data

Tests for normality, homogeneity of the data: Since we are assuming time series models in continuous time domain as well as conditional non-heteroskedasticity, for preliminary inspection, equal variance between samples in all data sets is tested [43]. To test this, we used the Levene's test for homogeneity [44] by dividing the data sets into subsets by the period of volatility (breaking the data sets suitably at the points where a local optimum is reversed) and any deviation in the variance between samples, thus constructed, is detected before we fit the models. If the resulting p-value arising from the test is less than some critical value (typically .05 - level of significance), it is inferred that the obtained differences in sample variances are unlikely to have occurred based on random sampling. The results from the Levene's test on the SPR and PV deaths data sets yield p-values of 0.6726 and 0.1743, respectively.

As a fundamental check of normality in the data sets, Q-Q plots and histograms are shown in Figure 1.

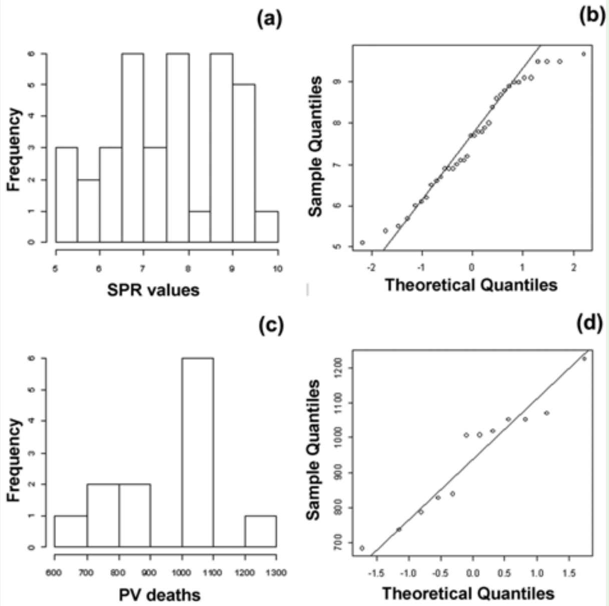

Figure 1: Histogram and Q-Q plots for: (a, b) SPR values, and (c, d) PV deaths.

To ascertain the existence of any seasonal pattern or cyclical behavior of the data sets we studied the Auto-Correlation Functions (Figure 2). Detailed discussion on ACF is available in Chatfield [41].

Models

Models used in our study are standard time series models. Brief descriptions of all these models are provided below:

AutoRegressive Integrated Moving Average models (ARIMA): The ARIMA models are the extension of the simpler AutoRegressive Moving Average (ARMA) models and are most well known class of models considered in time series studies, which helps to model the non-stationary time series data by converting them into a stationary series with appropriate differencing. We assume that historical data, comprising a time series $\{x_t: t=1,2,\ldots,n\}$ is provided, and then follow the Box-Jenkins methodology [45] to fit the model using the given time series. The major steps to be followed are: 1) Tentative identification of a model from the ARIMA class; 2) Estimation of parameters in the identified model; and 3) Diagnostic checks. If the tentatively identified model passes the diagnostic tests, the model is ready to be used for forecasting, if not, the diagnostic tests should indicate how the model ought to be modified, and a new cycle of identification, estimation and diagnosis is performed.

Fitting an ARMA (p,q) model: To fit an ARMA model, we look at either the correlogram or the partial correlogram and based upon the cutoff lag we decide an MA (q) or AR (p) model. If inference is difficult from cut off properties of correlograms, we look for an ARMA (p, q) model with non-zero values of p and q. We start with a simple model like ARMA (1,1) and if it is inadequate, we improve it to higher orders. Every additional parameter improves the fit of the model by reducing the residual sum of squares. But with increasing parameters and complexity of the model, the forecasts of the model may be misleading. Tetko *et al* [46] illustrates this as the problem of over-fitting, and hence there is a need for Akaike's Information Criterion (AIC), which is discussed later.

AutoRegressive Conditional Heteroskedastic models (ARCH): Like financial data [47], often in epidemiological time series the present value is influenced by the previous value (since previous disease incidence affects the future occurrence of the disease) and display a typical fat-tailed behavior (soon after a large change there exists a period of high volatility). This property of conditional variance is also known as conditional heteroskedasticity and is modeled using the ARCH class of models [37]. The major steps followed in these type of models are: 1) Specify the mean and volatility equation of the returns at time t; 2) For serial dependence in data, once the mean equation is specified the residuals of mean equation is used to test for ARCH effects and 3) Specify volatility model if ARCH effects are significant and perform a joint estimation of mean and volatility equations. An ARCH model with parameter p, defined as ARCH (p) is represented by the following equation,

$$X_t = \mu + e_t\sqrt{\alpha_0 + \sum_{k=1}^{p} \alpha_k (X_{t-k} - \mu)^2},$$

where$\{e_t\}$ is a sequence of i.i.d. standard normal random variables. Simplest representation of ARCH (p) is ARCH (1) defined as:

$$X_t = \mu + e_t\sqrt{\alpha_0 + \alpha_1 (X_{t-1} - \mu)^2}.$$

We have applied this modeling technique to model our data using X_t as the monthly SPR value / PV deaths.

Random Walk model: In data sets where the current values are dependent on the values at one lag in time [48], we often use Random Walk models satisfying the ITO process [49]. The general model is defined as:

$$X(t) = X(t-1) + \Delta Z(t),$$

Where $X(t)$ is the value of the series to be modeled and $\Delta Z(t)$ is the adjusted error term.

Recently, Chatterjee and Sarkar [28] demonstrated for the epidemiological data of the disease prevalence pattern of malaria is such that the Slide Positivity Rate (SPR) at the current time is highly

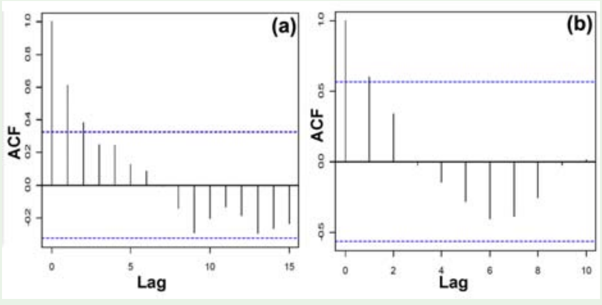

Figure 2: Auto-correlation functions of: a) SPR values and b) PV deaths.

Citation: Sarkar RR and Chatterjee C. Application of Different Time Series Models on Epidemiological Data - Comparison and Predictions for Malaria Prevalence. SM J Biometrics Biostat. 2017; 2(4): 1022.

correlated with slide positivity rate of the previous time point. The variation due to host population, vector population, migration effects of host and vector populations are few major causes apart from the other noise factors, which may distort the disease prevalence pattern as compared to the previous time point. To provide forecasts in a short forecasting horizon (for our response variables) it is possible to merge all these fluctuating factors into one term and can be accounted for the noisy fluctuations in and around the disease prevalence of the previous time point through the random walk.

We state the random walk model as the stochastic alternative of the auto-regressive model and define the model as:

$$S(t) = S(t-1) + dZ(t).$$

Where *S(t)* is the response variable (Monthly SPR/ Monthly PV deaths) at time *t* and *dZ(t)* is the error term defined as:

$$dZ(t) = \frac{(r(t) - \mu_r)}{\sigma_r},$$

Where, r(t) is the log of return of the random variable defined as:

$$r(t) = \ln(\frac{S(t)}{S(t-1)}).$$

It is to be noted that *E(dZ)=0* and *V(dZ)=1* for the random variable *dZ* reconfirms the fact that the ITO process is being satisfied. We checked that this model is equally adept in predicting short and swift fluctuations around the mean of stock price data as well as the more stable and factor-driven fluctuations in the case of malaria prevalence.

Diagnostic tests for model fitting

To assess the goodness of fit for the time series models, following diagnostics tests are carried out: (a) Augmented Dickey Fuller test to investigate the presence of a unit root, (b) Jarque-Bera test for normality of residuals and (c) Akaike's Information Criterion to determine the best fitted model. Brief descriptions for each of these tests are given below:

Augmented Dickey Fuller (ADF) test: The presence or absence of unit roots helps in identifying the characteristics of the underlying time series. This means, if a time series has no unit roots, it is stationary and will exhibit mean reverting behavior in long run and also will have a finite variance, independent of time. This is extremely crucial in forecasting, assuming that the random error decreases over time in absence of unit root. On the other hand in presence of a unit root, the process is non-stationary and exhibit very weak or no tendency to return back to a long run trend, along with possessing time dependent variance and tending towards infinity over a long horizon of time. Presence of a unit root in a sample time series is tested through the ADF test [50]. It is an augmented version of the Dickey Fuller test for larger and complicated time series. The following model is adopted initially

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \ldots + \delta_p \Delta y_{t-p} + \varepsilon_t,$$

where, α is a constant, β the regression coefficient for time trend and *p* the lag order of the autoregressive process. The unit root test is then carried out under the null hypothesis $\gamma = 0$ against the alternative hypothesis of $\gamma < 0$. Once a value for the test statistic is computed, it can be compared to the relevant critical value (DF) for the Dickey-Fuller Test:

$$DF = \frac{\hat{\gamma}}{SE(\hat{\gamma})}.$$

Jarque-Bera test for normality of residuals: Jarque and Bera [51] devised this test to compare the deviation of the assumed distribution with that of the normal distribution. The test is based on the sample kurtosis and skewness for the fitted model and it is the measure of departure from normality. The test statistic JB is defined as:

$$JB = \frac{n}{6}(S^2 + \frac{(K-3)^2}{4}),$$

Where *n* is the number of observations (or degrees of freedom in general); *S* is the sample skewness, *K* is the sample kurtosis and are defined as:

$$S = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{(\sigma^2)^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^{3/2}},$$

$$K = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{(\sigma^2)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}.$$

μ_3 and μ_4 are the third and fourth central moments, respectively, $\bar{x}$ is the samples mean, and σ^2 is the second central moment, the variance.

Akaike' s Information Criterion: One of the major objectives of this paper is to establish the best model that fits the epidemiological data. To establish this we have used the Akaike's Information Criterion, which measures the appropriateness of forecasts of the estimated statistical models [52] and acts as a better tool for model selection. Several models used in predicting for the same data, may be prioritized by their AIC measure. Increasing the number of free parameters in a model always improves the goodness of fit. AIC not only rewards goodness of fit, but also includes a penalty which is an increasing function of the number of estimated parameters and also attempts to find the model that best explains the data with minimum possible free parameters. It is imperative that the model with the lowest AIC measure will be regarded as the best.

In the general case, the AIC is given as *[-2ln(L)+2k]*, where '*k*' is the number of estimable parameters in the approximating model and '*L*' is the maximized value of the likelihood function for the estimated model. This '*k*' has also been referred as the asymptotic bias correction [53]. In the special case of Least Squares estimation with normally distributed errors,

$$AIC= 2k + n[ln(\hat{\sigma}^2)]$$

where $\hat{\sigma}^2$ is the residual variance of the predicted model.

Tools used in the analysis

The package R 2.9.0 is used extensively for modeling methods and generating test statistics. Binary packages developed by researchers for R 2.9.0 are used, namely (i) "TSA" [54] (ii) "tseries" [55] (iii) "sde" [56] (iv)"graphics" by R team [57] (v) "Rcmdr" [58] and (vi) "subselect" [59]. For other analytical and graphical analysis including simple linear regression, we used Microsoft Excel 2007 and SPSS for Windows 11.5 [60].

Results

In this study we compared the most frequently known models of the time series class, namely, ARIMA, GARCH, and the Random Walk, with respect to two epidemiological data sets of completely different nature, a longer time series (malaria prevalence in terms of SPR values) and one shorter time series (in terms of PV deaths).

For epidemiological data, we observe from the p-values (0.6726 for SPR and 0.1743 for PV deaths) of the Levene's test stated earlier, that we cannot reject the null hypothesis of homoskedasticity. This implies that for the assumption of constant variance, the epidemiological data sets will not behave erratically. After inspecting Q-Q plots and histogram (Figure 1) for these two types of data (SPR and PV deaths) it appears that the epidemiological data sets stray from normality. Moreover, the auto correlation functions of the SPR data set (Figure 2a) clearly indicate a seasonal pattern with the correlation function (exhibiting a sinusoidal curve over the number of lags). However, the PV deaths show no significant observable pattern (Figure 2b).

We tried several options of the general models as discussed in Section 2 and with all the available data sets we optimized the fit to the data with permutation of parameters of the general model. In the following subsections we discuss the outcomes of the model fitting for each of the data types.

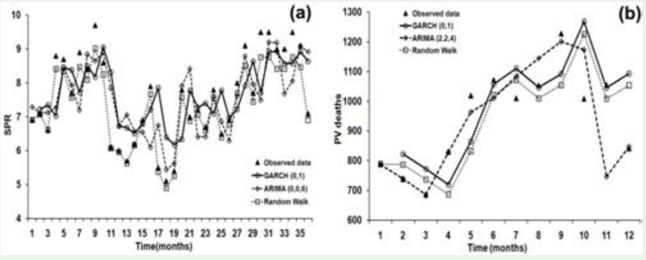

Figure 3: Fitted time series models for: a) SPR values and b) PV deaths.

Time Series Models

First to check the suitability of all these time series models we applied them on the stock price data obtained from an open source (the New York Stock Exchange [61]) and observed that all the models provide excellent fits (Supplementary Information). Then all these methods are applied to our epidemiological data and the corresponding models are obtained, which are summarized in the following sub-sections.

Models for Slide Positivity Rates of malaria prevalence: This epidemiological data is a long time series and is fitted with the same classes of models discussed before. Figure 3 (a) shows the comparative plot of fits for different time series models to the SPR values.

The characteristics for each of these fits are listed below and the statistical summary for the fitted time series is given in Table 1 and Table 2:

Table 1: Statistical summary for the diagnostic tests on the fitted time series models.

Model and diagnostics	Epidemiological data	
	SPR	PV deaths
ARIMA Model		
Jarque Bera Test		
Chi-square statistic	0.93	0.96
Degrees of freedom	2	2
p-value	0.63	0.62
Augmented Dickey-Fuller Test		
D-F statistic	-1.19	1.01
Lag order	3	2
p-value	0.89	0.99
ARCH Model		
Jarque Bera Test		
Chi-square statistic	0.53	0.62
Degrees of freedom	2	2
p-value	0.77	0.73
Augmented Dickey-Fuller Test		
D-F Statistic	-1.38	-0.73
Lag order	3	2
p-value	0.81	0.96
Random Walk Model		
Augmented Dickey-Fuller Test		
D-F statistic	-1.7	-1.46
Lag order	3	2
p-value	0.69	0.78
Jarque Bera Test		
Chi-squared statistic	1.93	0.65
Degrees of freedom	2	2
p-value	0.38	0.72

1. ARIMA (0,0,6): Moving Average model with dependency on 6 previous error points (the result justifies the pattern of the auto-correlation function shown in Figure 2(a)) is the best possible option among the ARIMA family. The yearly data fit generates errors that repeat half-yearly. This model gives R^2 as 52.22% (marked as $$, Table 2) signifying a moderately good fit.
2. The ARCH (0,1) is the best possible option from the GARCH family for fitting the SPR values and gives $R^2 = 34.27\%$ (marked as *#, Table 2), which is not very high compared to other fits. This is due to the fact that the SPR values in our data set do not show much variability in variance (also shown by the Levene's test as stated before). Thus the conditional variance is not highly volatile for the model to be related.
3. The Random walk model gives a high R^2 value, 98.6%, which shows that a noise term added to the previous time point SPR value is good enough to predict future values of SPR.

For the SPR series, we observe from Table 1 that on the basis of p-values of both JB and ADF tests we are unable to reject the null hypothesis. Based on these tests, we infer that the residuals of the fits for all the time series models for SPR are normal and the unit root of the predicted series for all the models are stationary. We note that for a fairly long time series of the epidemiological origin, the models provide good fits for SPR values, except the volatility cluster model, since logically this is not applicable for modeling this data set due to low variance differences between groups.

Models for *Plasmodium vivax* deaths*:* Figure 3(b) shows the comparative plot of fits for different time series models to *Plasmodium Vivax* (PV) deaths values. Like before the characteristics for each of these fits are listed below and the statistical summary for the fitted time series is given in Table 1 and Table 2:

1. ARIMA (2,2,4) is found to be the best model from the ARIMA family of models. This implies that the data series was not stationary in itself and differencing it two times makes the series stationary. Further on this differenced series we fit an ARMA model with 2 auto-regressive parameters and 4 error terms of previous time points. This gives an R^2 of 86.24%.
2. The ARCH (0,1) model gives the best fit from the GARCH family indicating general dependence of volatility terms on variances of previous series. This model has R^2 of 51.5% (marked as **, Table 2), which indicates not a good fit, possibly due to lack of volatile clusters that are heteroskedastic.
3. The Random Walk Model is not so successful as compared to the other fits that it shows in the other two cases. The reason is that this time series being highly fluctuating, has large deviations between the time points and cannot be harnessed in terms of noise that are accounted by the Random Walk model. The R^2 for this model is 45.69% (marked as ## in Table 2).

Table 2: Statistical summary for fitted time series models.

Models	Epidemiological data	
	SPR	PV deaths
ARIMA		
Model chosen	ARIMA (0,0,6)	ARIMA (2,2,4)
Model R^2	52.22% **$$**	86.24%
GARCH		
Model chosen	ARCH (0,1)	ARCH (0,1)
Model R^2	34.27% ***#**	51.5% ******
Random Walk		
Model R^2	98.60%	45.69% **##**

(different markers - $$, *#, **, ##" - are explained in the text for SPR and PV deaths)

For the PV death data, in all the models both JB and ADF tests generate non-significant p-values (Table 1) indicating that we are unable to reject the null hypotheses. We conclude that for our model-fits, the residuals are normal for all the models and the unit root is stationary for all predicted series of the data. In summary, in this case we observe that all models show promising result, except slightly low fitting of the random walk model. This happens due to failure in capturing the larger deviations between time differences, whereas the other models provide a fairly good fit.

Here we must mention that the paucity of available data is one of the major factors, especially the data set of deaths is very short (just 12 time points), which may run the risk of mis-fitting or over-fitting. But since we have used models which involve several parameters, we have used the Akaike criterion to sense over-fitting and have chosen the best available model with highest information and least number of possible parameters. With larger data sets and careful parameter choices one may be able to achieve even better results in terms of predicting the disease prevalence.

Forecasting and the best-fit model

To obtain the predictions based on the models that we applied to each of the data sets, we calculated the next few forecasted values from each time series models (Table 3) and compared with the available data. This clearly implies that on the basis of the given data, in a short forecasting horizon the models that we have chosen are very efficient in forecasting the future values. For all the models the predicted values are inside the 95% confidence limits of the observed values (shown within brackets in Table 3). It is worthy to note that in our study, the good fit of models are observed due to the fact that all the models used here have the inherent property to track the evolution of the previous values and predict the next. This conditional dependence is very obvious from the model formulae. Thus the models we chose may be adopted for modeling and forecasting purpose under suitable conditions.

Predicting large number of points for the PV deaths data set is actually unviable because in that case we run the risk of mis-fitting the data and basing a large number of predictions on a very short history. We therefore predicted only two time points for this data and compared it with respect to the observed values, which were available also only for two time points. But we have forecasted few more points for the longer time series data set of SPR (six points) and validated against the available observed data. In order to decide which model provides the best projections in the chosen horizon and to reinstate the suitability of the models for each of the data sets, we obtain the Akaike's Information Criterion for each of the model forecasts (Table 4).

Table 3: Predictions and comparisons of model's forecasting power for two types of malaria data.

Data sets	Observed (available data)	Model Predictions		
		ARIMA	*GARCH*	*Random Walk*
SPR ***(predictions for next 6 points)***	9.0 (8.21,9.84)	7.69	8.58	8.52
	9.5 (8.63,10.30)	8.06	8.58	8.86
	9.1 (8.29,9.93)	9.09	8.92	8.58
	7.1 (6.36,7.84)	8.92	8.65	6.99
	7.0 (6.25,7.72)	7.46	7.33	6.9
	**	7.31	7.46	##
PV deaths ***(predictions for next 2 points)***	1055 (271.05,1785.83)	746.31	1046.01	1009.27
	841 (202.28,1555.48)	848.23	1092.61	1053.4

** No data available for validation of the prediction

Prediction for this model is conditional on previous value

N.B. Values enclosed in braces are the 95% confidence limits for the best fitted model

For the Slide Positivity Rate data we observe that the AIC values (Table 4) suggest Random Walk model as the best model to provide better forecast. The ARCH model also provides a good approximation since there exists volatility clusters in the data. However due to parameter redundancy, the ARIMA models prove a failure, conforming the results observed from model fits. Moreover, the *Plasmodium vivax* deaths data also indicates that the Random Walk is the best model for this and ARCH model provides a very close second best fit (Table 4). ARIMA model fails to qualify the criterion, opposite to the result observed in the model fit.

Failure of a continuous time domain model in epidemiological data

To broaden our analysis apart from the three standard discrete time models, we tested the applicability of a continuous time domain model to see the differences in the model features in contrast to the discrete time behavior. For this purpose, we have chosen the Geometric Brownian Motion, which is the exponential form of the Brownian motion model and is the continuous analog of random walk model. This is a well-known model among the financial economists and has wide application since early twentieth century in asset price modeling [62]. We observe that it does not yield good results for the epidemiological data sets (Figure 4a and 4b), probably due to the fact that it is a diffusion process in continuous time domain with continuous state space. The R^2 and adjusted R^2 for SPR data set are 0.2% and -5.86% respectively and that for the PV deaths data set are 60.07% and 48.66% respectively.

Table 4: Akaike's Information Criterion (Â) for the model fits.

Data	Fitted Models		
SPR	**ARIMA (0,0,6)**	**GARCH (0,1)**	**Random Walk**
Sum of Squares	2083.05	2096.27	1974.51
Â	164.09	150.68	148.53
PV Deaths	**ARIMA (2,2,4)**	**GARCH (0,1)**	**Random Walk**
Sum of Squares	10855765.24	11074350.19	10902838.41
Â	206.58	170.16	169.97

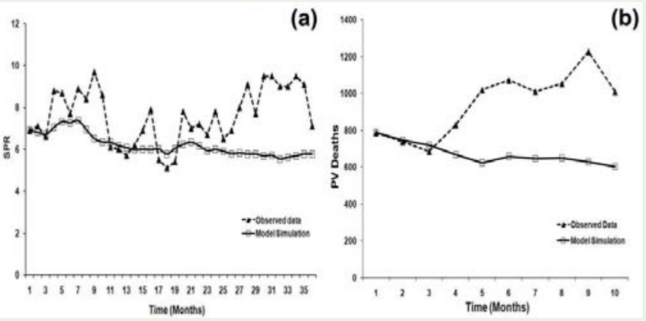

Figure 4: Simulated time series of the Geometrical Brownian motion for (a) SPR and (b) PV deaths.

A close observation to Figure 4(b) reveals that the fitted model and the observed data series fall apart and shows negatively correlated movement between the two series, which is also suggested by the R^2 values. However, from Figure 4 it is very clear that the model is not suitable for modeling epidemiological data, as demonstrated by our case data.

Discussion

Application of different Time Series model to analyse and predict financial data as well as epidemiological data is of long interest to researchers, even with existence of similarities and dissimilarities between the two data types. In both financial and epidemiological modeling, we are mainly faced with two types of times series data sets, broadly, stationary and non-stationary. The ARMA models are mainly used to model stationary time series data when the variance does not differ much between groups [41], whereas the ARCH class of models and their extension, the GARCH classes are used in data sets where volatility clusters are prevalent [38]. However in case the data exhibits non-stationarity, an extra parameter is added to ARMA to form an ARIMA model. Further, if we have a time series with noisy and short fluctuations such that it is difficult to differentiate between the trend and seasonal components, then we cannot adopt models of these types and use a Markov process, with an auto regressive term and an adjusted random noise component, widely known as Random Walk model [47,63].

Citation: Sarkar RR and Chatterjee C. Application of Different Time Series Models on Epidemiological Data - Comparison and Predictions for Malaria Prevalence. SM J Biometrics Biostat. 2017; 2(4): 1022.

The major objective of our work is to demonstrate that the models discussed above can be used successfully to analyse epidemiological data and also derive good forecast of disease prevalence. As is evident from the above study that the models that we have chosen are tailor made for the purpose of modeling stock prices as all the models namely the ARIMA family, the GARCH family, the Random Walk and the Geometric Brownian Motion (GBM) provide excellent fits for the stock price data set (Supplementary Information). However, the interesting observation is that these models also provide good fits when modeling Slide Positivity Rates and *Plasmodium vivax* deaths data with the only exception of GBM.

We observe from Figure 3 and Table 2, that the ARIMA class of models is always a good fit for the epidemiological data sets, which are mostly seasonal in nature. However, the ARCH model provides a fairly good estimation for the malaria data sets as well, though the fits are not highly encouraging and is the second best model according to the AIC metric for the three models taken together. This encourages us to believe that when we have stronger data sets with more durational gaps, we can use this model with high degree of confidence for modeling and forecasting of epidemiological time series data. Significantly, the Random Walk model provides a good model for the SPR data, as the fluctuations in the data set are very short and volatile. However, it seems to fail in case of data sets with longer seasonality and shorter length, as is the case with *Plasmodium vivax* deaths data. Even though the AIC metric (Table 4) shows that in this case also one can choose Random Walk model in contrast to other models, the large number of parameters reduce the model stability. We have shown that the models that we have chosen provide excellent fits and are good tools for forecasting epidemiological data in a short horizon. We have also shown in our study that with the use of these models one can predict the disease incidence and/or prevalence rates. For SPR we predicted 4 future time points and for the *Plasmodium vivax* deaths data we predicted 2 time points and observed that the Random Walk model gives excellent predictions. This forecasting can also be validated if the data is available (as we have shown for the SPR data set with our models). Depending on the nature, length and complexity of the epidemiological time series data one can choose appropriate models from these major classes and generate predictions accordingly.

In this study, we not only compared between data sets from financial and epidemiological background but also considered the applicability of models from different classes to each of these data so as to form a wider applicable ground for common class of models to compare, contrast and understand, and most importantly predict data series of different nature. The predicting and forecasting power exhibited by the chosen models will help researchers and public health professionals to develop disease control programmes and early warning systems for a disease prevalence forecast, specifically for malaria.

Acknowledgement

The authors thank Corporation of Chennai, Tamil Nadu, India for the malaria data. CC thanks Indian Academy of Sciences, Bangalore, for summer fellowship.

References

1. Bernoulli D. Essaid'une nouvelle analyse de la mortalite' cause'e par la petite ve'role et des advantages de l'inoculation pour la pre'venir. Me'm Math Physics Academy Royale Science Paris. 1760; 1-45.
2. Anderson RM, May RM. Infectious Diseases of Humans-Dynamics and control (1st edition). New York: Oxford University Press, 1991.
3. Kaufman J, Edlund S, Douglas J. Infectious disease modeling: Creating a community to respond to biological threats. Stat Comm in Infect Dis. 2009; 1: 1-14.
4. Mandal S, Sarkar RR, Sinha S. Mathematical models of malaria - a review. Malaria Journal, 2011;10: 202.
5. Bartlett MS. Measles periodicity and community size. J of Royal Statist Soc of America. 1957; 120: 48-70.
6. Wickwire K. Mathematical models for the control of pests and infectious diseases: a survey. Theor Pop Biol. 1977; 11: 182-238.
7. Agur Z, Cojocaru L, Mazor G, Anderson RM, Damon YL. Pulse Mass Measles Vaccination across Age Cohorts. Proceed of Natl AcadSci. 1993; 90: 11698-11702.
8. Rohani P, Earn DJD, Finkenstadtt B, GrenfellBT. Opposite Patterns of Synchrony in Sympatric Disease Metapopulations. Science. 1999; 286: 968-971.
9. Ludkovski M, Niemi J. Optimal Dynamic Policies for Influenza Management. Stat Comm in Infec Dis. 2010; 2.
10. Earnest A, Chen MI, Ng D, Sin LY. Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. BMC Health Ser Res. 2005; 11: 5-36.
11. Chang IS, Fu SJ, Chen CH, Wang TH, Hsiung CA. Estimating Temporal Transmission Parameters from Infectious Disease Household Data, with Application to Taiwan SARS Data. Stat Biosci. 2009; 1: 80-94.
12. Ture M, Kurt I. Comparison of four different time series methods to forecast hepatitis A virus infection. Exp Sys with Appln. 2006; 31: 41-46.
13. Dhand NK, Johnson WO, Toribio JLML. A Bayesian Approach to Estimate OJD Prevalence From Pooled Fecal Samples of Variable Pool Size. J of Agr Biol and Environ Stat. 2010; 15: 452-473.
14. Commenges D, Jolly D, Drylewicz J, Putter H, Thiébaut R. Inference in HIV dynamics models via hierarchical likelihood. Comp Stat and Data Anal. 2011; 55: 446-456.
15. World Health Organization (WHO). 2002.
16. Hay S, Guerra C, Tatem A, Noor A, Snow R. The global distribution and population at risk of malaria: past, present and future. The Lancet Infectious Diseases. 2004; 4: 327-336.
17. Muhammad K, Lenny B, Shunmay Y, Enny K, Noah W, Rilia M, et al. Malaria morbidity in Papua Indonesia, an area with multidrug resistant Plasmodium vivax and Plasmodium falciparum. Malaria Journal. 2008; 7: 148.
18. Chattopadhyay J, Sarkar RR, Chaki S, Bhattacharya S. Effects of environmental fluctuations on the occurrence of malignant malaria - a model based study. Ecol Model. 2004; 177: 179-192.
19. Ye Y, Louis VR, Simboro S, Sauebor R. Effect of meteorological factors on clinical malaria risk among children: An assessment using village-based meteorological stations and community-based parasitological survey. BMC Public Health. 2007; 7: 101.
20. Malaria risk in Africa (MARA) project.
21. Kuhn KG, Campbell L, Davies CR. A continental risk map of malaria mosquito (Diptera: Culicidae) vectors in Europe. J of Med Entom. 2002; 39: 621-630.

22. Kleinschmidt J, Bagayoko M, Clarke GPY, Craig M, Le D. A spatial statistical approach to malaria mapping. Sauer-International Epidemiological Association. 2000; 29: 355-361.

23. Ruru Y, Barrios EB. Poisson regression models of malaria incidence in Jayapura, Indonesia. The Philippine Statistician. 2003; 52: 27-38.

24. Briet O, Vounatsou P, Gunawardene DM, Galppaththy GNL, Amerasinghe PH. Models for short-term malaria prediction in Sri Lanka. Mal Journal. 2008; 7: 76.

25. Cancre N, Tall A, Rogier C, Faye J, Sarr O, Trape JF, et al. Bayesian analysis of an epidemiological model of plasmodium falciparum malarial infection in Ndiop, Senegal. Am J of Epidem. 2000; 152: 760-770.

26. Abellana R, Ascaso C, Aponte J, Saute F, Nhalungo D, Nhacolo A, Alonso P. Spatio-seasonal modeling of the incidence rate of malaria in Mozambique. Malaria Journal. 2008; 7, 228.

27. Gosoniu L, Vounatsou P, Sogoba N, Maire N, Smith T. Mapping malaria risk in West Africa using a Bayesian nonparametric non-stationary model. Comp Stat & Data Anal. 2009; 53: 3358-3371.

28. Chatterjee C, Sarkar RR. Multi-Step Polynomial Regression Method to Model and Forecast Malaria Incidence. PLoS ONE. 2009; 4: e4726.

29. Ito K. Stochastic integral. Proc of the Imper Acad of Tokyo. 1944; 20: 519-524.

30. Osborne MFM. Brownian motion in the stock market. Operations Research. 1959; 7: 145 - 173.

31. Boness AJ. Elements of a theory of stock-option value. The J of Pol Economy. 1964; 72: 163- 175.

32. Boyle PP. Options: A Monte Carlo Approach. J of Fin Econ. 1977; 4: 323 - 338.

33. Cox JC, Ross SA, Rubinstein M. Option Pricing: A simplified approach. J of Fin Econ. 1979; 7: 229-263.

34. King G, Charles IP, James HS, Mark WW. Stochastic Trends and Economic Fluctuations. The Am Econ Rev. 1991; 81: 819-840.

35. Hull JC, White A. The impact of default risk on options and other derivative securities. J of Bank and Fin. 1995; 19: 299-322.

36. Engle RF. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. Econometrica. 1982; 50: 987-1008.

37. Bollerslev T. Generalized autoregressive conditional heteroskedasticity. J of Econometrics. 1986; 31: 307-327.

38. Bollerslev T, Engle RF, Nelson DB. ARCH Models. Handbook of Econometrics. 1994; 4: 2959-3038.

39. Cohen I. Modeling speech signals in time-frequency domain using GARCH. Signal Processing. 2004; 84: 2453-2459.

40. Pérez A, Dennisa RJ, Rodríguezc B, Castroa AY, Delgadoc V, Lozanoa JM, et al. An interrupted time series analysis of parenteral antibiotic use in Colombia. J of Clin Epidem. 2003; 56: 1013-1020.

41. Chatfield C. The Analysis of Time Series-An Introduction (6th Edition). Florida: Chapman and Hall. 2004.

42. Holger K, Thomas S. Non linear Time Series Analysis (2nd edition) Edinburgh: Cambridge University Press. 2004.

43. Montgomery DC. Design and analysis of experiment. New York: Wiley. 1997.

44. Levene H. Robust tests for equality of variances in OLKIN, I. and ALTO, P. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. (1st edition). Stanford University Press. 278-292, 1960.

45. Box G, Jenkins G. Time series analysis: Forecasting and control (1st edition). San Francisco: Holden-Day. 1970.

46. Tetko IV, Livingstone DJ, Luik AI. Neural network studies 1. Comparison of Over fitting and Overtraining. J of Chem Inform and Comp Sci. 1995; 35: 826-833.

47. Tsay RS. Analysis of Financial Time Series (1st edition). Canada: John Wiley and Sons, 2002.

48. Dothan M. Efficiency and Arbitrage in Financial Markets. Intnl Res J of Fin and Econ. 2008; 19: 102-106.

49. Ito K. On stochastic differential equations. Memoirs Am Math Soc. 1951; 4: 1-51.

50. Dickey DA, Fuller WA. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. J of the Am Stat Assoc. 1979; 74: 427-431.

51. Jarque CM, Bera AK. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ Letters. 1980; 6: 255-259.

52. Akaike H. A new look at the statistical model identification. IEEE Trans on Automatic Contr. 1974; 19: 716-723.

53. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical-theoretic approach (2nd edition) New York: Springer-verlag. 2002.

54. Chan K. TSA: Time Series Analysis. R package version 0.97,2008.

55. Trapletti A, Hornik K. t series: Time Series Analysis and Computational Finance. R package version 0.10-18, 2009.

56. Iacus SM.sde: Simulation and Inference for Stochastic Differential Equations. R package version 2.0.7, 2009.

57. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

58. Fox J, Ash M, Boye T, Calza S, Chang A, Grosjean P, et al. Rcmdr: R Commander. R package version 1.4-10, 2009.

59. Silva JOCPD, Cadima J,Minhoto M.subselect: Selecting variable subsets. R package version 0.9-9993, 2009.

60. SPSS Inc.

61. The New York Stock Exchange. Available: http://www.nyse.com/listed/ibm.html

62. Bachelier L. Scientific Annals of the École Normale Supérieure.. (English translation- A. J. Boness, Cootner P H. (1964) The random character of stock market prices. Cambridge, MA: MIT Press. 1900; 17 - 75.

63. Samuel K, Howard MT. A first course in stochastic processes (2nd edition). New York: Academic Press. 1975.

Citation: Sarkar RR and Chatterjee C. Application of Different Time Series Models on Epidemiological Data - Comparison and Predictions for Malaria Prevalence. SM J Biometrics Biostat. 2017; 2(4): 1022.