British Journal of Mathematics & Computer Science
4(3): 375-393, 2014

SCIENCEDOMAIN *international*
www.sciencedomain.org

Mathematical Modeling of the Epidemiology of Tuberculosis in the Ashanti Region of Ghana

Kwame Asare Gyasi-Agyei[1*], Amoakoh Gyasi-Agyei[2] and William Obeng-Denteh[3]

[1]Department of Applied Mathematics, Koforidua Polytechnic, P.O. Box 981, Koforidua, Ghana.
[2]Ghana ICT Research Institute (GICTRAC), P. O. Box AH163, Accra, Ghana.
[3]Department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Original Research Article

Received: 27 June 2013
Accepted: 20 September 2013
Published: 07 November 2013

Abstract

Tuberculosis is an infectious disease which can be fatal. Hence, availability of models predicting its potential outbreak can be very useful in its preventative strategies. This paper finds the best mathematical model which fits onto the tuberculosis occurrence data of Ashanti Region of Ghana, and uses the model to predict the future epidemiology and incidence of the disease in the region to enhance anti-tuberculosis campaigns. The data used for the study was obtained from the Ashanti Health Services and spans January 2001 to March 2013. It is evident from the analysis that tuberculosis occurrence in the region studied can best be modeled with ARMA (1, 0) or AR(1), i.e. a stochastic time series linear model, and that tuberculosis epidemic in the Ashanti Region is expected to be stable between April 2013 and April 2015. The Mean Absolute Error (MAE) and the Mean Squared Error (MSE) are used to compare the in-sample forecasting performance of three selected competing models, and the result shows that it is not always true that the best selected model gives the best results so far as the mean square error (MSE) is concerned. The forecasting accuracies for the obtained model, i.e. AR (1), using MAE and MSE are respectively 16.3171 and 461.3148.

Keywords: *ARMA Model in Health; Medical Statistics; Tuberculosis Forecasting; Tuberculosis Modeling;*

1 Introduction

An estimated 14 million people worldwide are infected with active tuberculosis. However, the vast majority of tuberculosis-related deaths occur in the developing world. In 2009 alone, 9.4 million new tuberculosis cases were reported, and 380,000 out of 1.7 million reported deaths were related to tuberculosis among people with HIV [1]. TB in Ghana predates the country's independence but the then government recognized the need to combat the disease due to the threat it posed to the larger society [2]. In 2011, 8.7 million people suffered from TB-related illnesses,

**Corresponding author: gaakus@yahoo.com, gyasi-agyei@gictrac.org, obengdentew@yahoo.com;*

including 1.1 million cases among people with HIV,leading to 1.4 million deaths [3]. The good news is that, according to the report of World Health Organization (WHO) in November 2010, the number of new TB cases continues to fall in five of the six WHO regions. The exception is Southeast Asia, where incidence remains stable. Worldwide deaths relating to tuberculosis fell by 35 percent between 1990 and 2009 [1]. Thus, on the average, TB prevalence is on the decline worldwide.

The epidemiology of tuberculosis varies substantially around the world. The highest rates (100/100,000 or higher) are observed in sub-Saharan Africa, India, China, and the islands of Southeast Asia and Micronesia. Estimates provided by USAID in 2007 for South Sudan were 228 cases per 100,000 population. In South Sudan, an estimated 18,500 people develop TB, leading to 5,300 TB-related deaths annually [4]. TB is one of the top killers of women worldwide; half a million women died from TB in 2011. The TB mortality rate has decreased by 41% since 1990. In 2011, 5.8 million newly diagnosed cases worldwide were recorded in national TB control programmes (NTPs). This is still only two-thirds of the estimated total of 8.7 million people who were diagnosed of TB in that year. In that same year 2011, 48% of the TB patients known to be living with HIV globally were put on antiretroviral therapy (ART) [3]. Implementation of this collaborative anti-TB/HIV campaign saved an estimated 1.3 million lives globally between 2005 and 2011.

TB is a disease of poverty affecting mainly young adults in their most productive years [1].TB is an infectious disease that is caused by a bacterium called *Mycobacterium tuberculosis*[23], which is a member of the Mycobacterium complex. The other members in this complex are *Mycobacterium africanum* and *Mycobacterium bovis*. *Mycobacterium africanum* is most commonly found in West Africa; it causes up to a quarter of cases of tuberculosis in the Gambia [4].TB primarily affects the lungs, but it can also affect other parts of the body such as the central nervous system, lymphatic system, kidney, spine, brain and the circulatory system. The disease was called "consumption" in the past because of the way it consumes from within anyone who is infected with it [5].

If not treated properly, tuberculosis can be fatal [6]. *Mycobacterium bovis* is the main cause of tuberculosis in cattle, deer, and other mammals. The human bacillus might have arisen from *M. bovis* in the setting of animal domestication. Human *M. bovis* infection generally occurs in the setting of consumption of infected cow's milk products, Bacille Calmette-Guérin (BCG) vaccination for TB prevention, or intravesicular BCG installation for bladder cancer treatment [4]. The primary stage of TB does not cause symptoms. Symptoms of pulmonary TB can include: cough that lasts three weeks or longer, presence of blood or sputum in cough, pain in the chest, loss of weight, loss of energy, poor appetite, fever, chills, excessive sweating especially at night, fatigue, fever, weight loss, breathing difficulty, and wheezing [11], [7].Symptoms often improve in 2 to 3 weeks after starting treatment.

TB is a contagious disease that spreads through the air. When people with the disease cough, sneeze, talk or spit, they propel TB germs, known as *bacilli*, into the air. Only a small number of the bacilli need to be inhaled to cause an infection [8]. Babies and young children often have weak immune system and thus much more vulnerable to the infection. Other vulnerable people are those with any of these conditions: HIV infection, substance abuse, silicosis (8), diabetes mellitus, severe kidney disease, low body weight, head and neck cancer, medical treatments such as corticosteroids or organ transplant, specialized treatment for rheumatoid arthritis or Cohn's disease [4]. However, not all people infected with TB bacilli feel sick. The immune system either

kills the germs, or "walls off" the TB bacilli where they can lie dormant for years. Failure of the immune system to control infection with TB bacilli leads to active disease, when TB bacilli multiply and cause damage in the body.

TB can be recognized through medical tests such as tuberculin skin tests, chest x-ray, and sputum smear tests. People infected with either latent or active TB usually show a positive tuberculin skin test and blood test results. Only people with active TB produce positive sputum smear results or have an abnormal chest x-ray [6]. The following are the tests for pulmonary TB: Biopsy of the affected tissue, Bronschoscopy, Chest CT scan, Chest X-ray, Interferon-gamma release blood test such as the QFT-Gold test to test for TB infection, Sputum examination and cultures, Thoracentesis, Tuberculosis skin test [7].A chest x-ray usually does not show improvement from treatment until weeks or months later. Outlook is excellent if pulmonary TB is diagnosed early and effective treatment is started quickly. People are advised to seek medical treatment if they get exposed to TB, develop symptoms of TB, symptoms continue despite treatment, or new symptoms develop [4].

Left untreated, each person with infectious TB can spread the germs to about 10 to 15 people every year [8]. Medications are the cornerstone of tuberculosis treatment. However, treating TB takes much longer than treating other types of bacterial infections. With tuberculosis, the patient must take antibiotics for at least six to nine months. The exact drugs and length of treatment depend on the patient's age, overall health, possible drug resistance, the form of TB and the infection's location in the body [9]. The goal of treatment is to cure the infection with drugs that fight the TB bacteria. Treatment of active pulmonary TB always involve a combination of many drugs. All of the drugs are continued until laboratory tests show which medicines work best. Commonly used drugs include: Isoniazid, Rifampin (Rifadin, Rimactane), Pyrazinamide and Ethambutol (Myambutol). Other drugs that may be used to treat TB include: *Amikacin*, *Ethionamide*, *Moxifloxacin*, *Para-aminosalicylic* acid and *Strptomycin* [7]. There is some evidence that taking vitamin D during tuberculosis treatment enhances some of the effects of the drugs [3].

A recent study suggests that a shorter term of treatment — three instead of nine months — with combined medication may be effective in keeping latent TB from becoming active TB. With the shorter course of treatment, people are more likely to take all their medications and the risk of side effects is lessened. More study is needed to ascertain this recent finding. A TB patient may need to take just one type of TB drug. Active tuberculosis, particularly if it is a drug-resistant strain, will require several drugs at once [8].

Side effects of TB drugs are not common but can be serious when they do occur. All tuberculosis medications can be highly toxic to the liver. Patients on such medications are advised to inform their doctors immediately if they experience any of the following: nausea or vomiting, loss of appetite, a yellowish skin color (i.e. jaundice), dark urine, a fever that lasts three or more days and has no obvious cause [9]. Pulmonary TB can cause permanent lung damage if not treated early. Other side effects of anti-TB drugs include: changes in vision, orange- or brown-colored tears and urine, or rash. A vision test may be done before treatment so doctors can monitor any changes in the health of the eyes [7].

Persons who have been recently infected with TB bacteria include: persons who have close contacts with a person with infectious TB disease, persons who have immigrated from areas of the world with high rates of TB infections , children less than 5 years of age who have a positive TB

test, persons from groups with high rates of TB transmission (such as homeless persons, injection drug users, and persons with HIV infection), and persons who work or reside with people who are at high risk of TB in facilities or institutions such as hospitals, homeless shelters, correctional facilities, nursing homes and residential homes for those with HIV [4].

Pulmonary TB is preventable, even in those who have been exposed to an infected person. Skin testing for TB is used in high risk populations or in people who may have been exposed to TB, such as health care workers. People who have been exposed to TB should be skin tested immediately and have a follow-up test at a later date, if the first test is negative. A positive skin test means there has been a contact with the TB bacteria. Such people need to consult a medical doctor on how to prevent getting tuberculosis. Prompt treatment is extremely important in preventing the spread of TB from those who have active TB disease to those who have never been infected with TB. Some countries, such as Ghana, with a high risk of TB infections vaccinate people with BCG vaccine to prevent TB. However, the effectiveness of this vaccine is limited and it is not routinely used in some countries including the United States. People who have had BCG may still be skin tested for TB. If the test results positive then it must be discussed with a medical doctor [7].

From the study of the open literature no research work has been so far reported on the model of epidemiology of tuberculosis in the Ashanti Region of Ghana. This paper therefore seeks to fill the research gap using time series analysis. The main objective of this paper is to study the pattern of tuberculosis infections in the Ashanti Region. The remaining part of this paper is organized as follows. Section 1.1 surveys the research work relating to the subject of the paper. Section 2 discusses the data used, how they were obtained as well as the methods applied on them to obtain the needed results. Section 3 deals with empirical analysis of the paper. Section 4 contains the concluding remarks.

1.1 Literature Review

From the work of [10], despite the infectious agent that causes tuberculosis having been discovered in 1882, many aspects of the natural history and transmission dynamics of TB were still not fully understood. This was reflected in differences in the structures of mathematical models of TB, which in turn produced differences in the predicted impacts of interventions. Gaining a greater understanding of TB transmission dynamics required further empirical laboratory and field work, mathematical modelling and interaction between them. Modelling could be used to quantify uncertainty due to different gaps in their knowledge to help identify research priorities. Fortunately, the present moment was an exciting time for TB epidemiology, with rapid progress being made in applying new mathematical modelling techniques, new tools for TB diagnosis and genetic analysis and a growing interest in developing more-effective public-health interventions.

Reference [11] addresses the spread of tuberculosis through one-strain and two-strain models. They first presented a basic model that incorporated fast and slow progression, effective chemoprophylaxis and therapeutic treatment. The system exhibited the traditional behavior. They proved that if the basic reproduction ratio R_0 = 1, then the disease-free equilibrium is globally asymptotically stable on the nonnegative out-hunts. However, if R_0>1 an endemic equilibrium exists and is globally asymptotically stable. Based on the first model, the second model dealt with

the problem of drug resistance as a competition between multiple types of strains of mycobacterium tuberculosis: those that were sensitive to anti-tuberculosis drugs and those that were resistant. Their objective was to characterize the role of multi-drug-resistant in the transmission of tuberculosis. The coexistence and stability of the associated equilibria were discussed.

According to [12], TB was a leading cause of infectious mortality. Although anti-biotic treatment was available and there was vaccine, tuberculosis levels were rising in many areas of the world. They used mathematical models to study tuberculosis in the past and had influenced public policy. The spread of HIV and the emergence of drug-resistant TB strains motivated the use of mathematical models today. Here, they reviewed and compared the mathematical models of tuberculosis dynamics in the literature. They presented two models of their own: a spatial stochastic individual-based model and a set of delay differential equations encapsulating the same biological assumptions. They compared two different assumptions about partial immunity and explored the effect of preventative treatments. They argued that seemingly subtle differences in model assumptions could have significant effects on biological conclusions.

[13] on their part evaluated the efficacy of recommended tuberculosis infection control measures. They used a deterministic mathematical model for airborne contagion. They examined the percentage of purified protein derivative conversions under various exposure conditions, environmental control strategies, and respiratory protective devices. They concluded that environmental control cannot eliminate the risk for TB transmission during high-risk procedures. However, respiratory protective devices, and particularly high-efficiency particulate air masks, may provide nearly complete protection if used with air filtration or ultraviolet irradiation. Nevertheless, the efficiency of those control measures decreased as the infectivity of the source case increased. Therefore, administrative control measures were the most effective because they substantially reduced the rate of infection.

Reference [14] studied the spread of tuberculosis through a two-patch epidemiological system $SE_1 \cdots E_nI$ which incorporated migrations from one patch to another just by susceptible individuals. Their model was considered with bilinear incidence and migration between two patches, where infected and infectious individuals cannot migrate from one patch to another due to medical reasons. They discussed the existence and uniqueness of the associated endemic equilibria. They used quadratic forms and Lyapunov functions to show that when the basic reproduction ratio is less than one, the disease-free equilibrium (DFE) is globally asymptotically stable, and when it is greater than one there exists in each case a unique endemic equilibrium which was globally asymptotically stable. Numerical simulation results were provided to illustrate the theoretical results.

Reference [15] formulated mathematical models to establish the conditions on the size of the area occupied required minimizing and thereafter eradicating tuberculosis. Both numerical and qualitative analyses of the model were done and the effect of variation in the area size and recruitment rate on the different epidemiological groups was investigated. Their results of the analysis showed that there exists a stable disease-free equilibrium point provided that the characteristic area was greater than the product of the probability of survival from the latent stage to the infectious stage and the number of latent infections produced by a typical infectious individual during his/her mean infectious period. Their study recommended that the characteristic area per individual should be at least 0.25 square kilometres in order to minimize the tuberculosis incidence.

Reference [16] reviewed a mathematical model of the dynamical behaviour of tuberculosis disease in the Upper East Region of the Northern part of Ghana. The equilibrium points of the model system were found and their stability was investigated. His model exhibited two equilibria, namely, the disease-free equilibrium and the endemic equilibrium. He used stability theory and computer simulation, and observed that population determine the infection rate of tuberculosis hence the higher the population density, the greater the risk of instability of the disease-free equilibrium.

The use of different mathematical tools to study biological processes is necessary to capture effects occurring at different scales. They studied as an example the immune response to infection with the bacteria Mycobacterium tuberculosis, the causative agent of tuberculosis. They showed that the immune responses were both global as well as local in nature. They used four different mathematical tools to explore both the global immune response as well as the more local one and compared and contrasted results obtained using those methods. Applying a range of approaches from continuous deterministic models to discrete stochastic ones allowed them to make predictions and suggested hypotheses about the underlying biology that might otherwise go unnoticed. The tools developed and applied were also applicable in other settings such as tumor modeling [17].

The strengths and limitations of using homogeneous mixing and heterogeneous mixing epidemic models were explored in the context of the transmission dynamics of tuberculosis in [18]. Their focus was on three types of models: a standard incidence homogeneous mixing model, a non-homogeneous mixing model that incorporates 'household' contacts, and an age-structured model. The models were parameterized using demographic and epidemiological data and the patterns generated from those models were compared. Furthermore, the effects of population growth, stochasticity, clustering of contacts, and age structure on disease dynamics were explored. That framework was used to assess the possible causes for the observed historical decline of tuberculosis notifications.

The study looked at the estimation of the economic burden and household welfare impact of tuberculosis (TB) in the Western Region of Ghana. Studies into the economic burden of TB in Ghana have been limited. WHO‘s (2002a) guidelines on cost and cost effectiveness of TB management were followed in the estimation of cost of TB from the patient/household and health provider perspectives. Human capital method was applied in the cost estimation. Wells-Riley model and multiple regression technique were employed in the estimation of the probability of transmission within households and the household welfare impact of TB. Results established that tuberculosis causes a significant deterioration in household income and welfare. The study also found that TB imposes various catastrophic economic costs on affected households and utilize considerable resources within the public health system. It is recommended that safety nets or income insurance be establish for households affected by TB to help them cope with high economic burden as well as helping patients fully complete treatment [19].

According to [20], tuberculosis (TB) is a serious public health issue in developing countries. Early prediction of TB epidemic is very important for its control and intervention. They aimed to develop an appropriate model for predicting TB epidemics and analyze its seasonality in China. Data of monthly TB incidence cases from January 2005 to December 2011 were obtained from the Ministry of Health, China. They used a seasonal autoregressive integrated moving average (SARIMA) model and a hybrid model which combined the SARIMA model and a generalized regression neural network model to fit the collected data from 2005 to 2010. Simulation

performance parameters of mean square error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to compare the goodness-of-fit between these two models. Data from 2011 TB incidence data was used to validate the chosen model. From their results, although both two models could reasonably forecast the incidence of TB, the hybrid model demonstrated better goodness-of-fit than the SARIMA model. For the hybrid model, the MSE, MAE and MAPE were 38969150, 3406.593 and 0.030, respectively. For the SARIMA model, the corresponding figures were 161835310, 8781.971 and 0.076, respectively. The seasonal trend of TB incidence is predicted to have lower monthly incidence in January and February and higher incidence from March to June. They concluded that the hybrid model showed better TB incidence forecasting than the SARIMA model. There is an obvious seasonal trend of TB incidence in China that differed from other countries.

2. Materials and Methods

The data used for the modeling and analysis was obtained from the Ministry of Health in the Ashanti Region of Ghana. It consists of monthly tuberculosis cases from various hospitals in the Ashanti Region for the period of January 2001 to March 2013.

Ashanti region is used as our case study area since the region is a cosmopolitan region in Ghana. All sorts of people are living in the region. The region is the third largest of 10 administrative regions in Ghana, occupying a total land surface of 24,389 square kilometers or 10.2 per cent of the total land area of Ghana. In terms of population, however, it is the most populated region with a population of 4,780,380 in 2010 population and housing census (PHC), accounting for 19.4 per cent of Ghana's total population; however, its density (148.1 per square km) is lower than those of the Greater Accra (895.5/km^2) and Central (162.2/km^2) Region [21]. The region is centrally located in the middle belt of Ghana. It lies between longitudes 0.15W and 2.25W, and latitudes 5.50N and 7.46N. The region shares boundaries with four of the ten political regions, Brong-Ahafo Region in the north, Eastern region in the east, Central region in the south and Western region in the south west. The region is divided into 30 districts, each headed by a district chief executive [22]. Majority of the region's population are Ghanaians by birth (87.3%) with about five per cent naturalized Ghanaians. A smaller proportion (5.8%) of the population originate from outside Ghana, made up of 3.7 per cent mainly from the five English-speaking countries of ECOWAS and 2.1 per cent from other African countries. The non-African population living in the region is 1.8 per cent of the total population. Akans are the predominant ethnic group in the region, representing 77.9% of Ghanaians by birth. A high proportion (78.9%) of the Akan population is Asante. The non-Akan population in the region comprises the Mole- Dagbon (9.0%), the Ewe (3.2%), the Grusi (2.4%), the Mande-Busanga (1.8%) and the Ga- Dangme (1.4%). The other smaller ethnic groups form about 1.3 per cent of the population of the region [23].

The basis of the Box-Jenkins approach to modeling time series is summarized below and consists of three phases: identification, estimation and testing, and application [24].

2.1 Identification Stage

In this identification stage, we transform the collected data to be used for the modeling into stationary data with stabilized variance. For non-seasonal data, first differencing is usually sufficient to attain stationarity. The first-order difference of a time series is defined as

$$\nabla X_t = X_t - X_{t-1}, \tag{1}$$

where X_t is a random variable at time *t*. Another very important diagnostic tool for examining interdependence of data is the sample autocorrelation functions. In this section, we examine the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the data to identify potential models.

The autocorrelation function (ACF), $\rho_{t,s}$, is defined as

$$\rho_{t,s} = Corr(X_t, X_s) = \frac{Cov(X_t, X_s)}{\sqrt{Var(X_t)Var(X_s)}}, \ \forall\, t, s \in \{0, \pm1, \pm2, \pm3, \dots\}, \tag{2}$$

where $Cov(X_t, X_s) = E[(X_t - \mu_t)(X_s - \mu_t)] = E(X_t, X_s) - \mu_t\mu_s$, μ_t is the expected value of the stochastic process (X_t) at time *t*.

The sample autocorrelation function, r_k, at lag *k* is given by equation (3), where $\overline{X}$ is the grand mean.

$$r_k = \frac{\sum_{t=k+1}^{n}(X_t - \overline{X})(X_{t-k} - \overline{X})}{\sum_{t=1}^{n}(X_t - \overline{X})^2}, \text{ for } k = 1, 2, 3, \dots. \tag{3}$$

For a variety of reasons, this has become the standard definition for the sample autocorrelation function. A plot of r_k versus lag *k* is often called a *correlogram* [25].

2.2 Estimation and Testing Stage

2.2.1 The Estimation Phase

We now proceed with the general development of autoregressive moving average (ARMA), and mixed autoregressive moving average (ARMA) models for the stationary time series obtained in

Section 2.1. A time series $\{X_t; t= 0, \pm1, \pm2, \pm3, \ldots\}$ is ARMA (p, q) if it is stationary and fulfills equation (4).

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \ldots + \theta_q \omega_{t-q} + \omega_t. \quad (4)$$

$\phi_p \neq 0$, $\theta_q \neq 0$ and $\rho_\omega{}^2 > 0$. The parameters $p$and $q$are called the autoregressive and the moving average orders, respectively. In the situation in which X_t has a non-zero mean μ, we set $\alpha = \mu(1 - \phi_1 - \phi_2 - \ldots - \phi_p)$ and use the alternative model

$$X_t = \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \ldots + \theta_q \omega_{t-q} + \omega_t, \quad (5)$$

where $\{\omega_t; t= 0, \pm1, \pm2, \ldots\}$ is a Gaussian white noise sequence and α is the intercept [26]. An example of an ARMA (p, q) model is the ARMA (1, 1), which is defined as equation (6).

$$X_t = \alpha + \phi_1 X_{t-1} + \theta_1 \omega_{t-1} + \omega_t. \quad (6)$$

The ARMA (1, 1) model is stationary if $-1 < \phi_1 < 1$ and it is invertible if $-1 < \theta_1 < 1$. In an ARMA (1, 1) model both ACF and the PACF trail off to zero.

2.2.2 Testing stage

While the approximate linear decay of the sample ACF is often taken as a symptom that the underlying time series is nonstationary and requires differencing, it is also useful to quantify the evidence of non-stationarity in the data-generating mechanism. This can be done via hypothesis testing.

Consider the model $X_t = \phi_1 X_{t-1} + Y_t$ for t = 1, 2, 3, …,where $\{Y_t\}$ is a stationary process. The process $\{X_t\}$ is non-stationary if the coefficient ϕ_1 = 1, but it is stationary if $|\phi_1| < 1$. Thus, the null hypothesis corresponds to the case where the AR characteristic polynomial has a unit root and the alternate hypothesis states that it has no unit roots [26].

2.2.2.1 Augmented dickey-fuller (ADF) test

The hypothesis H_0: X_t is non–stationary against the alternate hypothesis H_1: X_t is stationary can be tested in the regression equation (7).

$$\Delta X_t = \beta_0 + \alpha t + \beta_1 X_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta X_{t-i} + \varepsilon_t. \quad (7)$$

Accept H_0 if $P-value > 0.05$, else accept H_1.

2.2.2.2 Kwiatkowski-phillips-schmidt-shin (KPSS) test

An alternative approach to the ADF test is the KPSS test. A hypotheses of H_0: X_t is level or trend stationary is tested against H_1: X_t is non- stationary in the regression equation (8).

$$X_t = \alpha_t + \beta t + \mu_t, \quad (8)$$

where a random walk, $\alpha_t = \alpha_{t-1} + \varepsilon_t$ is allowed. We accept H_0 if $P-value > 0.05$, else we accept H_1.

2.2.2.3 Testing the model for adequacy (portmanteau test)

After identifying an appropriate model for a time series data, it is very important to check that the model is adequate. Reference [27] provides a modified portmanteau test statistic for checking the randomness of the error terms. Their statistic is given by equation (9).

$$Q^* = n(n+2) \times \sum_{k=1}^{h} \left(\frac{{r_k}^2}{n-k} \right). \quad (9)$$

Q^* is approximately distributed as a χ^2 with $h-p-q$ degrees of freedom, where n is the length of the time series, h is the first h autocorrelations being checked, p is the order of the auto-regressive process, q is the order of the moving average process, and r_k is the estimated autocorrelation coefficient of the k^{th} residual term. If the calculated value of Q^* is greater than χ^2 for $h-p-q$ degrees of freedom, then the model is considered inadequate and the model is adequate if Q^* calculated is less than χ^2 for $h-p-q$ degrees of freedom. If the model is tested inadequate, then the forecaster should select an alternative model and test for the adequacy of the model [24].

2.3 Application: Selecting Best Model Using Suitable Criterion

A number of other approaches to model specification have been proposed since Box and Jenkins' seminal work. One of the most studied is Akaike's Information Criterion (AIC). This criterion advocates to selecting the model that minimizes AIC [25]. The AIC is equal to twice the number of parameters in the model minus twice the logarithm of the likelihood function. Mathematically, AIC is calculated as

$$AIC(p,q) = 2k - 2\log(\text{maximum likelihood}), \tag{10}$$

where $k = p + q + 1$ if the model contains an intercept or constant term, and $k = p + q$ otherwise.

Given two or more competing models, the one with the smallest AIC value is deemed more appropriate. The AIC is a biased estimator, and the bias can be appreciable for large parameter per data ratios. Reference [28] showed that the bias can be approximately eliminated by adding another non-stochastic penalty term to the AIC, resulting in the corrected AIC, denoted by AICc and defined by equation (11).

$$AICc = AIC + \frac{2(k+1)(k+2)}{n-k-2}. \tag{11}$$

n is the sample size and k is the total number of parameters excluding the noise variance.

Another approach to determining the ARMA orders is to select a model that minimizes the Schwarz Bayesian Information Criterion (BIC). It is defined mathematically by equation (12).

$$BIC(p,q) = k \times \log(n) - 2 \times \log(\text{maximum likelihood}). \tag{12}$$

If the true process follows an ARMA (p, q) model, then it is known that the orders specified by minimizing the BIC are consistent; that is, they approach the true orders as the sample size increases. However, if the true process is not a finite-order ARMA process, then minimizing AIC among an increasingly large class of ARMA models enjoys the appealing property that it will lead to an optimal ARMA model that is closest to the true process among the class of models under study [25].

We now turn to another fundamental concern, i.e. how to measure the suitability of a particular forecasting method for a given data set. In most forecasting situations, accuracy is treated as the overriding criterion for selecting a forecasting method. In many instances, the word "accuracy" refers to "goodness of fit," which in turn refers to how the forecasting model is able to reproduce the data that are already known. To the consumer of forecasts, it is the accuracy of the future forecast that is most important [24]. Among the common used to measure estimation accuracy are the mean absolute error (MAE) and the mean squared error (MSE).

The MAE and the MSE are defined respectively as

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t|. \tag{13}$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}e_t^{\,2}. \tag{14}$$

X_t is the actual observation of the process at time *t*, F_t is the forecast value of the process at the same time, $e_t = X_t - F_t$ is the error term, and *n* is the number values of the process estimated.

3. Data Analysis and Results

This section discusses the analysis and modeling of the data collected on tuberculosis cases in the Ashanti Region of Ghana. Here, we applied the **R** Statistical Package in modeling the time series. That is, all the plots and numerical output displayed in this paper have been produced with the **R** software. Most of the numerical outputs have been edited for additional clarity or for simplicity. Actual tuberculosis data drawn from various hospitals in the Ashanti Region of Ghana are used throughout in this paper to illustrate the methodology presented in Section 2.

3.1 Time Plot of Prevalence of Tuberculosis (TB) in the Ashanti Region of Ghana

In general, the trend in TB prevalence in the Ashanti Region of Ghana seems to be irregular. The annual TB time plot in Fig. 1 does not exhibit seasonal variation. The green line shows the mean of the series. Most of the data points are very close to the mean. This indicates that there is a clear case of stationarity in the mean. It follows that the TB series is stationary in the mean since the data fluctuate around a constant mean, independent of time, and the variance of the fluctuation remains essentially constant over time. There was not any seasonal behavior in the time plot, and hence the TB data looks to be approximately stable for further investigations.

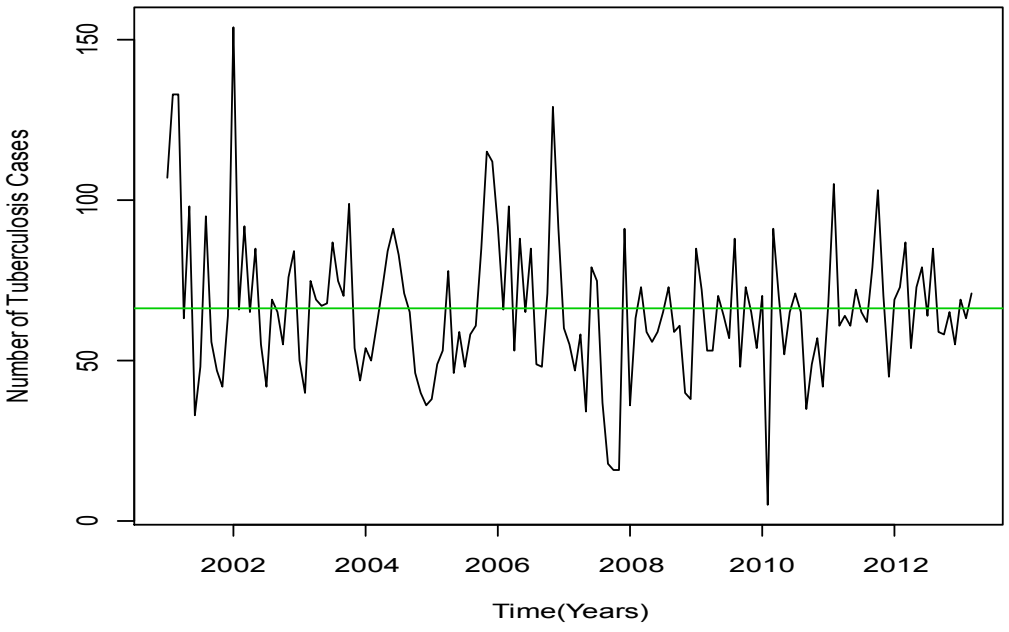

Fig. 1. TB cases in the Ashanti-Region of Ghana from January 2001 to March 2013

3.1.1 Stationarity checks using the ACF and PACF

Fig. 2 depicts the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the TB data. From Fig. 2 (i) the spikes from 2 to 20 autocorrelation do not exceed two standard errors above zero (they are significantly near to zero). This shows that there is a stationarity in the TB data. Fig. 2 (ii) exhibits the partial autocorrelation function (PACF) of the TB data. At lag zero the PACF is far away from unity (1) which confirms that the tuberculosis time series is stationary.

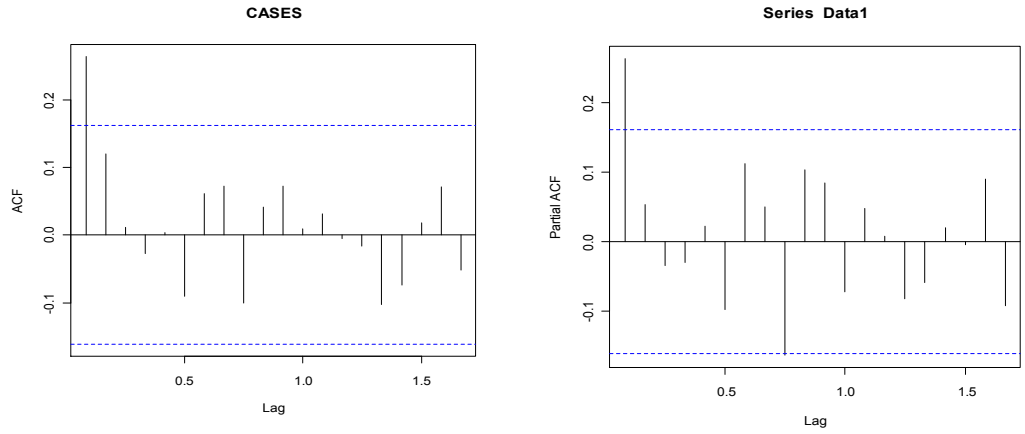

Fig. 2 (i) **Fig. 2 (ii)**

Fig. 2. The Autocorrelation Function and Partial Auto-Correlation Function of TB prevalence

3.1.2 Stationarity checks using the augmented Dickey-Fuller (ADF) Test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

The Augmented Dickey-Fuller (ADF) Test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test for tuberculosis data are discussed in Table 1. The table shows both ADF and KPSS test results. The P -value indicates the stationarity or otherwise of the time series under study. The time series is stationary if and only if the P-value for the ADF test falls below 0.05 (i.e. 5% significant level) and that of KPSS exceeds 0.05. As these conditions are fulfilled, we can conclude that the original tuberculosis (TB) data of the Ashanti Region of Ghana is stationary. The two tests confirmed that there was stationarity in the original TB data and hence mean stationarity is achieved.

Table 1. ADF and KPSS Tests

ADF Test		
Dickey-Fuller	**Lag Order**	**P-value**
-5.1017	5	0.01
KPSS Test		
KPSS Level	**Lag Parameter**	**P -value**
0.2961	2	0.1

3.2 Selecting Competing Models Using ACF and PACF of TB Cases

The first part of Fig. 2 shows the sample ACF of the TB cases. Except for marginal significance at spike 1, the model seems to have captured the essence of the dependence in the series. Inspecting the sample ACF, we see that PACF is tailing off and the ACF is cutting off at spike 1 as shown in Fig. 2 (i). This suggests that the TB data follows an MA (1) model.

Fig. 2 (ii) shows the sample PACF of the TB cases in the Ashanti Region of Ghana at different lags. Inspecting the sample PACF, we see that the ACF is tailing off and the PACF is cutting off at spike 1. Except for marginal significance at spike 1, the model seems to have captured the essence of the dependence in the series. This suggests an AR (1) for the TB cases.

As a preliminary analysis, we will fit both models. It follows that, in both the ACF and the PACF of the tuberculosis (TB) data in Fig. 2, the following models were suggested:

1. ARMA (0, 1) or MA (1);
2. ARMA (1, 0) or AR (1)
3. ARMA (1, 1).

3.3 Estimation of Tentative Models

The estimations of the three selected tentative models with non-zero mean are discussed below.

3.3.1 Parameter estimate and diagnostics of ARMA (0, 1) model

Table 2 depicts the parameter estimate for ARMA (0, 1) with non-zero mean. The coefficient of the estimated MA(1) parameter is within the causality condition bounds since its absolute t value is greater than 2 as shown in Table 2. From Table 2, the estimated ARMA (0, 1) model can be written as shown in equation (15)

$$X_t = 0.2194\,\omega_{t-1} + 71.5024\,. \tag{15}$$

Table 2. Parameter Estimate for ARMA (0, 1) with Non-zero Mean

Coefficient	Estimate	Standard Error	$\lvert t$ – value$\rvert$	Intercept
ma1	0.2194	0.0735	2.9850	71.5757
AIC	***AICc***	***BIC***		***Constant***
1326.41	1326.69	1333.37		-0.0733

Results from Table 3 showed that the model's residuals were non - significant with Ljung Box test statistic of 14.8703 and a P-value of 0.7838. Hence the model was adequate for forecasting.

Table 3. Box-Ljung test of ARMA (0, 1) with non- zero mean

X-Squared	Degrees of Freedom (df)	P -value
14.8703	20	0.7838

3.3.2 Parameter estimate and diagnostics of ARMA (1, 0) model

The coefficient of the estimated AR (1) parameter is within the causality condition bounds since its absolute *t*- value is greater than 2 as shown in Table 4. From Table 4 the estimated ARMA (1, 0) model can be written as

$$X_t = 0.2557\ X_{t-1} + 71.6667\ . \tag{16}$$

Table 4. Parameter estimate for ARMA (1, 0) with non-zero mean

Coefficient	Estimate	Standard error	|*t* – value|	Intercept
ar1	0.2557	0.0802	3.1883	71.7416
AIC	***AICc***	***BIC***		***Constant***
1324.95	1325.23	1336.91		-0.0749

Results from Table 5 shows that the model's residuals were non-significant with Ljung Box test statistic of 15.2002 and a *P*-value of 0.7648. Hence the model was adequate for forecasting.

Table 5. Box-Ljung test and Forecasts from ARMA (1, 0) with non- zero mean

X-Squared	Degrees of freedom (df)	P -value
15.2002	20	0.7648

3.3.3 Parameter estimate and diagnostics of ARMA (1, 1) model

From Table 6 the coefficients of the estimated ARMA (1, 1) parameters are outside the causality and invertibility condition bounds since their absolute *t*-values are less than 2. From Table 6 the estimated ARMA (1, 1) model can be written as shown in equation (17).

$$X_t = 0.4184X_{t-1} - 0.1716\omega_{t-1} + 71.8699. \tag{17}$$

Table 6. Parameter estimate for ARMA (1, 0) with non-zero mean

Coefficient	Estimate	Standard error	|*t* – value|	Intercept
ar1	0.4184	0.2537	1.6492	71.9467
ma1	-0.1716	0.2687	0.6386	
AIC	***AICc***	***BIC***		***Constant***
1326.57	1326.99	1341.52		-0.0768

Results from Table 7 showed that the model's residuals were non - significant with Ljung Box test statistic of 15.2481 and a *P*-value of 0.7620. Thus the model was adequate for forecasting.

Table 7. Box-Ljung test and Forecasts from ARMA (1, 1) with non- zero mean

X-Squared	Degrees of Freedom (df)	P -value
15.2481	20	0.7620

3.4 Forecasting from ARMA (1, 0), the Best Model

From Fig. 3, the yellow region depicts the 95% confidence interval, the red region is the 85% confidence interval and the blue line is the forecasting points. The model was used to forecast two years ahead and showed that the tuberculosis prevalence in the Ashanti Region of Ghana will be stable from April 2013 to April 2015 as shown in the blue line.

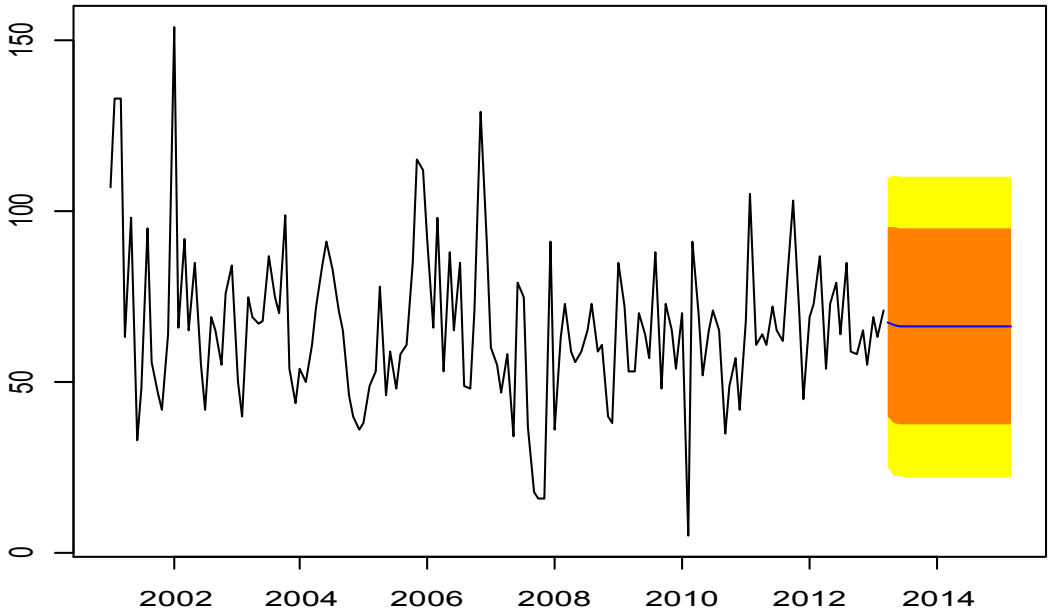

Fig. 3. Forecasts from ARMA (1, 0) with non-zero mean

3.5 The Error Metrics

Forecasting accuracy based on the Mean Absolute Error (MAE) of the forecasted values was checked for each fitted model as shown in Table 8. It highly favored the forecasted value of ARMA (1, 0), the best selected model. This means that, the ARMA (1, 0) forecast error of 16.3171 out-performed all the forecast errors so far as the MAE is concerned. Hence ARMA (1, 0) was confirmed to be the best model.

Similarly, the forecasting accuracy based on the Mean Squared Error (MSE) of the forecasted values favored ARMA (1, 1), one of the competing selected models. This means that, the ARMA (1, 1) forecast error of 460.1114 out-performed all the forecast errors so far as the MSE is concerned.

Table 8. The Mean Absolute Error (MAE) and the Mean Squared Error (MSE)

(i) Mean Absolute Error (MAE)		(ii) Mean Squared Error (MSE)	
ARMA(0, 1)	16.3394	ARMA(0,1)	465.9801
ARMA(1, 0)	16.3171	ARMA(1, 0)	461.3148
ARMA(1, 1)	16.3172	ARMA(1, 1)	460.1114

4. Conclusion and Recommendations

Results from Table 9 compares three of the best models tested for the TB data. It can be observed that ARMA (1, 0) shows the lowest AIC value of 1324.96. It is thus the best model for the TB data so far as the AIC was concerned.

Table 9. Summary of Diagnostics Test

MODEL	AIC	AICc
ARMA (1, 0)	1324.96	1325.23
ARMA (0, 1)	1326.35	1326.69
ARMA (1, 1)	1326.57	1326.99

Hence the best model for the TB data is $X_t = 0.2557\, X_{t-1} + 71.6667$. However, the forecasting accuracy based on the MAE for ARMA (1, 0), the best obtained model, was the lowest and its forecast error was calculated as 16.3171. In conclusion then, the research study reported in this article has found that tuberculosis data in the Ashanti Region of Ghana is best modeled with ARMA (1, 0) or AR(1). The study again found out that tuberculosis prevalence in the Ashanti Region is expected to be stable from April, 2013 to April 2015.

This model did not consider mass vaccination as one of the methods to prevent the prevalence of tuberculosis in the region. The results of this paper can be used as a tool to facilitate the introduction of tuberculosis vaccine and improve tuberculosis vaccination in the country as a whole. The results of this paper showed that it is not always true that the best selected model gives the best results so far as the mean squared error (MSE) is concerned. From above discussions, we conclude that the tuberculosis data from the Ashanti Region of Ghana spanning January 2001 to March 2013 depicts a s*tochastic time series* with a *linear* model, i.e. ARMA (1,0) or AR (1). Thus, for any chance of eradicating the disease, it is recommended that Ghana Health Service may not relent on its mass tuberculosis vaccination campaign in the region.

Competing interests

Authors have declared that no competing interests exist.

References

[1] Global Fund (2010), www.theglobalfund.org/Tuberculosis.

[2] Ghana Health Service. Health Programmes; 2012.
Available: http://www.ghanahealthservice.org/health_prog.php.

[3] World Health Organization. Global Tuberculosis Report; 2012. Available on line at: www.who.int/**tb**/publications/factsheet_global.pdf.

[4] Robert L, Serafino Wani. Epidemiology of mycobacterium tuberculosis, Specialist Trainee in Infectious diseases & Medical Microbiology/Virology, Royal Free Hospital, London, UK; 2012. Available on line at www.southsudanmedicaljournal.com/archive/may-2012/...

[5] Crosta Peter MA. Medical News Today; 2009.
Available: www.medicalnewstoday.com/articles/8856.php, 0:00 PST

[6] Canadian Centre for Occupational Health and Safety; 2012.
Available: www.cchs.ca/oshanswers/disease/tubercul.html, OSH

[7] National Center for Biotechnology Information.U.S. National Library of Medicine,8600 Rockville Pike, BethesdaMD, 20894USA; 2012.
Available on line at: www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001141.

[8] World Health Organization; 2009. Available: www.who.int/features/qa/o8/en/index.html

[9] Mayo Clinic; 2012. www.mayoclinic.com/.../**tuberculosis**/.../DSECTION=**treatments**-and-drugs.

[10] White PJ, Garnett GP. Mathematica lmodelling of the epidemiology of tuberculosis.AdvExp Med Biol. 2010;673:127- 40. Available on line at: www.ncbi.nlm.nih.gov/pubmed/20632534.

[11] Claude Jean Kamgang, Bowong Samuel, Jean Jules Tewa. Stability analysis of the transmission dynamics of tuberculosis models;2011.
www.worldacademicunion.com/journal/1746.../wjmsvol07no02paper01.pdf

[12] Caroline C, Ted C, Megan M. Mathematical Models of Tuberculosis: Accomplishments and Future Challenges;2006. Proceedings Trim Size: 9in x 6in colijn˙, seis.bris.ac.uk/~encgc/colijn_brazil.pdf

[13] Laura Gammaitoni, Maria Clara Nucci. Using a Mathematical Model to Evaluate the Efficacy of TB Control Measures;1997. wwwnc.cdc.gov/eid/article/3/3/97-0310_article.htm

[14] Jean Jules Tewa, Samuel Bowong, Oukouomi Noutchie CS. Mathematical analysis of two-patch model of tuberculosis disease with staged progression, Applied Mathematical Modelling. 2012;36:5792-5807; doi:10.1016/j.apm.2012.026

[15] Sematimba A, Mugisha JYT, Luboobi LS. Mathematical Models for the Dynamics of Tuberculosis in Density-Dependent Populations: The Case of Internally Displaced Peoples' Camps (IDPCs) in Uganda, Academic journal article from Journal of Mathematics and Statistics.2005;1(3).

[16] Adetunde IA. The Mathematical Models of the Dynamical Behaviour of Tuberculosis Disease in the Upper East Region of the Northern Part of Ghana: A Case Study of Bawku. Current Research in Tuberculosis. 2009;1:15-20.DOI:10.3923/crt.2009.15.20, URL: http://scialert.net/abstract/?doi=crt.2009.15.20

[17] Gammack David, Suman Ganguli, Simeone Marino, Jose Segovia-juarez, Denise,KirschnerE. Understanding the immune response in tuberculosis using different mathematical models and biological scales;2005.

[18] Aparicio Juan Pablo, Carlos Castillo-Chavez Abstract. Mathematical modelling of tuberculosis epidemics, Mathematical biosciences and engineering MBE. Publisher: School of Science and Technology, Universidad Metropolitana, San Juan 00928-1150, Puerto Rico. juan.p.aparicio@gmail.com.2009;6(2):209-237.

[19] Blankson Hammond Kodwo. Economic Burden of Tuberculosis in the Western Region of Ghana; 2012.

[20] Cao S, Wang F, Tam W, Tse LA, Kim JH, Liu J, Lu Z. A hybrid seasonal prediction model for tuberculosis incidence in China, BMC Medical Information Decision Making. 2013;13:56. doi: 10.1186/1472-6947-13-56.

[21] Ghana Statistical Service. Population and Housing Census (PHC); 2010.

[22] Wikimedia Commons; 2013.http://commons.wikimedia.org/wiki/File:Ashanti_districts.png

[23] Wikipedia; 2013., http://en.wikipedia.org/wiki/Ashanti_Region

[24] Spyros M, Wheelwright SC, Hyndman RJ. Forecasting: Methods and Applications, Third Edition, U.S.A., John Wiley & Sons, Inc; 1998.

[25] Cryer D. Jonathan, Kung– Sik Chan. Time Series Analysis With Applications in R, Second Edition, U.S.A., Springer; 2008.

[26] Shumway HR, Stoffer SD. Time Series Analysis and Its Applications With R Examples, Second Edition, U.S.A., Springer; 2006.

[27] Box GEP, Jenkins GM, Reinsel GC. Time Series Analysis: Forecasting and Control, 3rd edition, Prentice Hall, Englewood Cliffs, New Jersey; 1994.

[28] Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika.1989;76(2):297–307.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid= 313&id=6&aid=2447