

Cancer Cases in Kenya; Forecasting Incidents Using Box & Jenkins Arima Model

Amos Langat¹, George Orwa¹, Joel Koima²

¹Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

²Department of Mathematics and Informatics, Kabarak University, Nakuru, Kenya

Email address:

moskiplangat@gmail.com (A. Langat), orwa@fsc.jkuat.ac.ke (G. Orwa), jkoima@kabarak.ac.ke (J. Koima)

To cite this article:

Amos Langat, George Orwa, Joel Koima. Cancer Cases in Kenya; Forecasting Incidents Using Box & Jenkins Arima Model. *Biomedical Statistics and Informatics*. Vol. 2, No. 2, 2017, pp. 37-48. doi: 10.11648/j.bsi.20170202.11

Received: December 13, 2016; **Accepted:** January 17, 2017; **Published:** February 15, 2017

Abstract: The aim of the study was to fit appropriate time series models in assessing the accuracy of the Box Jenkins and ARIMA model in forecasting of Cancer case admissions for all people of any age from different health facilities across the country. Box-Jenkins was selected for evaluation because it has the potential of producing a point forecast within a given population, it provides a forecast interval, and is based upon a proven model. Forecast results and their associated forecast intervals may help Health facilities and health practitioners make informed decisions about whether the number of observed cancer reports in a given timeframe represents a potential incidence or is a function of random variation. Data management and analysis were done in SPSS Software. The data was segmented into two sets: Training Set (from 2000 to 2015) and the Test Set (from 2016 to 2018). The hold out set (test) provides the gold standard for measuring the model's true prediction error which refers to how well the model forecasts for new data. To note, the test data were only be used after a definitive model has been selected. This was to ensure unbiased estimates of the true forecast error. The results were presented in form of tables, graphs and context. In this study, the developed model for cancer case incidents in Kenya was found to be an ARIMA (2,1,0). From the forecast available by using the developed model, it can be seen that forecasted incidents for the year 2015-16 is higher than 2014-15 and in later years the incidents increases. The model can be used by researchers for forecasting of cancer incidents in Kenya.

Keywords: Box-Jenkins, ARIMA Models, Forecasting, Cancer Incidence

1. Introduction

Cancer has been one of the major causes of morbidity and mortality globally, but by a bigger extend in the African Region [1]. Cancers are caused by combined genetic and non-genetic changes induced by environmental factors that trigger inappropriate activation of specific genes leading to neoplastic transformations or abnormal cell growth. There is lack of information about key cellular events that occur in early stages of cancer development as well as environment factors and internal cause that trigger these changes.

Advances in molecular epidemiology are allowing researchers the possibility of simultaneously identifying multiple changes affecting the genome and extra-genomic should be now possible to define which genetic and other alteration or combinations thereof, can be interpreted as reliable biomarkers of exposures [2]. Cancer differs from

most other diseases in that it can developed at any stage in life and in any body organ. No two cancer cases behave exactly alike; some may follow an aggressive course, with the cancer growing rapidly. Very high cure rates can be achieved for some types of cancers, but for others the cure rates are disappointingly low and await improved for methods of detection and treatment. The most common stage of Cancer at diagnosis, the rate of progression, and the treatment option vary significantly with the type of Cancer a patient present [3].

It is estimated that about 80% of Cancer are due to environment or lifestyle, and therefore are potentially preventable [2]. The risk factors for some Cancers have been clearly identified but for others further research are needed, Based on current evidence, at least 30% of future Cancer cases are preventable by comprehensive and carefully considered action taken now [4].

The cancer treatment that a patient receives is determined

by the stage of Cancer at diagnosis, the type and location of the Cancer, the standard medical practical practices and treatment guidelines in the patient's County [5] and the ability of the Patient to pay for treatment for most standard and effective form of initial Cancer treatment. Multiple metastases (in various locations) and the overall tumours load ultimately limit surgical removal and the effectiveness of anti-cancer drugs. When cancers recur and spread beyond the initial site or region, systemic treatment is necessary and the goal of this treatment is no longer curative. Chemotherapy is the most prevalent form of systemic treatment, because it can reach and destroy cancer cells throughout the body, although the blood-brain barrier often limits effectiveness in the case of brain metastases. Chemotherapy may be used alone or in combination with other forms of treatment such as radiation therapy to specific metastatic sites. Hormone-regulated tumours, such as certain breast and prostate cancers use the body's natural hormones to grow, and they are often more responsive to hormone-based treatments than chemotherapy. As in the case of chemotherapy, tumours can become increasingly resistant to standard treatments. Certain cancers can be resistant to systemic treatments at the time of diagnosis. Overall, 30% to 80% of cancers can become refractory [4, 5, 6].

In Kenya, Cancer ranks third as a cause of death after infectious diseases and cardiovascular diseases. In Kenya, the risk of getting Cancer before the age of 75 years is 14% while the risk of dying of Cancer is estimated at 12%. In many developing countries the rapid rise in Cancer and other non-communicable diseases has resulted from increased exposure to risk factors which include tobacco use, harmful use of alcohol and exposure to environmental changes. Other risk factors for some Cancers include infection diseases such as HIV/AIDS, Human Papilloma Virus, Hepatitis B and C (Liver Cancer), bacterial infections such as *Helicobacter pylori* (Cancer of Stomach) and parasitic infestations such as schistosomiasis (Cancer of bladder).

The leading Cancer in women are breast Oesophagus and cervical Cancers. In men, Oesophagus and prostate Cancer and keposi sarcoma are the most common cancers. According to regional Cancer registry at KEMRI, about 80% of reported cases of Cancer are diagnosed at advanced stages, when very little can be achieved in terms of curative treatment. This largely due to the low awareness of cancer signs and symptoms, inadequate screening services, inadequate diagnostic facilities and poorly structured referral facilities. This makes it difficult for a great majority of the population to access cancer treatment services resulting in long waiting times causing some previously curable tumours to progress to incurable stage.

The reason for this situation is that, cancer treatment infrastructure in Kenya is inadequate and some cancer management options are not readily available necessitating some Kenyans to seek cancer treatment abroad. Some of the very essential drugs for pain management are rare to find in most public hospitals where majority of population can easily to access.

There is also comprehensive cancer surveillance system and no population based cancer registry [10]

In Kenya, the easily available and accessible sources of data for Cancer indicators which involve population-based surveys such as Demographic Health Surveys (DHS) are burdened with many limitations including data incompleteness, inaccuracies and are unavailable in actionable timeframes. Health facility-based data has been an alternative to these cross-sectional surveys due to its longitudinal nature and hence has been applied in many scenarios to determine short and long-term disease trends. However, health facility data should be utilized with the full recognition of its limitations [11]. The application of various mathematical and statistical methods on the cancer incidence will be instrumental and enabling the modelling of disease trends over time, with additional methods that account for the shortcomings of the available data. Cancer is a leading cause of death worldwide. It is projected that an estimated 15.5 million people will be diagnosed, and 12 million will die of cancer in the year 2030. The annual mortality attributed to main types of cancer includes: lung cancer (1.3 million deaths), stomach cancer (803 000 deaths), colorectal cancer (639 000 deaths), liver cancer (610 000 deaths) breast cancer (519 000 deaths) cervical cancer (450,000) and esophageal cancer (380,000). The most frequent types of cancer among men affect the lung, stomach, liver, colorectal, Oesophagus and prostate. Among women the most common areas affected are breast, lung, stomach, colorectal and cervix [12, 13].

The annual incidence of cancer globally is estimated to be 10 million. Of these, 5.3 million are in developing countries. In developed countries, cancer is the second most common cause of death after cardiovascular conditions and epidemiological evidence points to the emergence of a similar trend in developing countries. The principal factors contributing to this projected increase in cancer are the increasing proportion of elderly people in the world (in whom cancer occurs more frequently than in the young), an overall decrease in deaths from communicable diseases, the decline in some countries in mortality from cardiovascular diseases, and the rising incidence of certain forms of cancer, notably lung cancer resulting from tobacco use. Approximately 20 million people are alive with cancer at present, and by 2020 this number is projected to increase to more than 30 million [14].

Regardless of prognosis, the initial diagnosis of cancer is perceived as a life-threatening event, with over one-third of patients experiencing clinical anxiety and depression. Cancer is also distressing for the family, profoundly affecting both the family's daily functioning and economic situation. The economic shock includes both the loss of income and the expenses associated with health care costs. The strategy aims to build on the existing health system in Kenya to strengthen cancer and control capacities both in public and private sectors through control of risk factors associated with cancer, investment in cancer control workforce, equipment and through cancer research. This is the first cancer control. It consolidates aspects in cancer prevention, screening,

diagnosis, treatment and care for cancer patients as well as investment needed to deliver these services.

The strategy particularly reinforces the need for action to prevent cancer, especially those related to smoking and other modifiable risk factors. Enhanced health promotion, education and advocacy will enable the government and other partners to improve public understanding of cancer. It will empower the public general, to adopt healthier lifestyles and healthcare professionals in particular to recognize the symptoms of cancer and identify people at risk or living with cancer. It seek to improve early detection of cancer by introducing or expanding the available screening programmes and putting in place mechanisms and services that are proven to lives. It seek to shorten the time taken to diagnose and treat cancer by streamlining the diagnosis and referral systems, the process of care and investing in more cancer treatment equipment as well as cancer specialist and other staff. The strategy also seek to improve access to cancer drugs and other aspects of care for cancer patients. This strategy seek to harmonize and coordinate cancer care, national cancer registration, sharing of resources and information among health facilities. It will ensure patients and their families have better support and access to quality treatment including palliative care. In the field of cancer prevention and control, it will ensure a culture of evidence based practice [15].

Health facility-based data has been an alternative to these cross-sectional surveys due to its longitudinal nature and hence has been applied in many scenarios to determine short and long-term disease trends. However, health facility data should be utilized with the full recognition of its limitations [16]. The application of various mathematical and statistical methods on this data has been instrumental and has enabled the modelling of disease trends over time, with additional methods that account for the shortcomings of the available data.

The global community and country-specific efforts led to the adoption of various interventions to curb the disease burden, efforts which have been sustained and improved over time [17].

However, a direct causality link between interventions and trends over time has not yet been established which could be due to the unavailability of consistent data to measure indicators, the possibility of intervention supply information redundancy from different stakeholders.

There are three major categories of forecasting methods which include; qualitative/judgmental methods, quantitative methods and technological methods. Qualitative forecasting methods are generally subjective in nature and are based on expert opinions in formulating relationships. Conversely, quantitative forecasting methods involve statistical procedures in analysis of past values or historical data to establish true mathematical relationships or approximate associations which are reasonably closer to the truth [18, 19]. The three sub-categories of quantitative methods include; time series methods which seek to identify historical patterns using time as a reference point and then forecast future

values using time-based extrapolation procedures; explanatory methods which seek to identify past relationships producing observed outcomes and forecasting by applying the established relationships in the future; and monitoring methods which seek to identify the changes in relationships and patterns. Technological methods address societal, economic, political or technological long-term issues using expert-based methods, or historical relationships, patterns and analogies to define and “forecast” pre-determined future values [20].

There are a number of factors to consider before selecting the appropriate technique to use in forecasting. Some of these factors include; the type of data being analyzed, the time horizon which can be classified as either short, medium or long term; the level of detail or frequency required which increases with need; the number of series' and parameters involved and the historical patterns and constancy [19]. When data is presented with time as the reference point, time series models are often adopted. These are unique methods whose usage first involves the understanding of the structure of an observed set of data and the underlying forces producing it through estimation and fitting of appropriate models, and second, using these established relationships for forecasting, monitoring and evaluation, or even feedback and feedforward control. The time series approach has often been used to model disease patterns, explaining what will happen but not why [21].

Reliable, consistent and timely data is a prerequisite for effective planning and implementation of Cancer prevention, control and case management interventions. The major shortcomings hindering this process in many low and middle income countries include; poor vital registration systems, lack of comprehensive Health Management Information Systems (HMIS) and disease surveillance systems, unavailability of longitudinal data over time and if available, of low quality, incomplete and/or inaccessible in actionable timeframes [33]. For instance, Cancer is a leading cause of death worldwide; According to the 2008 World Cancer Report shows that the global burden of Cancer doubled between 1975 and 2000 and is expected to doubled again by 2020 and nearly triple by 2030 [34].

To accommodate this, statistical modelling has been deployed often to create tools which are able to obtain more accurate estimates for the required data which are closer to the truth. The use of this method has evolved over time, from merely studying underlying forces and structures that produced particular sets of observed data to modelling and forecasting into the future [35].

Box-Jenkins and ARIMA model was selected for evaluation because it has the potential of producing a point forecast within a given population, it provides a forecast interval, and is based upon a proven model [36]. Forecast results and their associated forecast intervals may help Health facilities and health practitioners make informed decisions about whether the number of observed disease reports in a given timeframe represents a potential incidence or is a function of random variation. This is possible as Box-Jenkins

ARIMA relies upon the mathematical properties of the underlying time series from which the forecast is based and not upon the dynamics of infectious disease transmission. Box-Jenkins is an autoregressive integrated moving average (ARIMA) model. The difference between traditional regression and ARIMA is that the variable being forecast is not related to another variable but is related to its own past values, a process known as autocorrelation [37]. Autocorrelation examines the correlation between each observation and its previous observations. Moreover, forecasts based upon ARIMA take into account the premise that data, taken over time, may have an internal structure based upon trend, and seasonality that can be accounted for [38]. Many diseases exhibit trend and seasonality [39], and as such, the use of Box-Jenkins ARIMA was an appropriate tool to make these forecasts.

The ability in predicting Cancer incidence accurately is a major milestone in the control and management of the disease. The results obtained could facilitate optimal distribution of resources, enabling the adoption of appropriate control interventions tailor made to the county, region or transmission setting under consideration. This will in the long run lead to the reduction in the number of new and resurgence Cancer cases and Cancer-attributable deaths on the path to fulfilling the global and country-specific targets for the disease at large.

2. Materials and Methods

This researcher used health facility-based data obtained from Five Hospitals located in different part of the Country which include; Narok Hospital (Narok County), Longisa Hospital (located in Bomet County), Tenwek Mission Hospital (located in Bomet County), Kapkatet Hospital (located in Kericho County), and AIC Litein Hospital (located in Kericho County). The primary data on case admissions for all cancers patients were collected over time from health facilities records for the period 2000-2015.

Data management and analysis were done in SPSS Software. The data were segmented into two sets: Training Set (from 2000 to 2015) and the Test Set (from 2016 to 2018). The hold out set (test) provides the gold standard for measuring the model's true prediction error which refers to how well the model forecasts for new data. To note, the test data were only be used after a definitive model has been selected. This was to ensure unbiased estimates of the true forecast error. The results were presented in form of tables, graphs and context.

This research focuses in detail on the understanding of the Box-Jenkins ARIMA Modeling Approach and the Time Series Methods used in the analysis and forecasting of Cancer incidence case admissions. Methodologies used in modelling, and features incorporated into each model were some of the main aspects of the section. Forecasting methods can be broken into time series and explanatory types of analysis [42]. Time series models tend themselves to predicting the continuation of historical patterns, such as

disease burden within a community, if the three following conditions are met:

1. Data from the historical record are available.
2. These data are quantified in the form of numerical data.
3. It can be assumed that at least some portion of the past pattern will continue into the future.

2.1. The Model

The Box-Jenkins approach refers to a set of procedures for identifying, estimating, checking and even forecasting a time series model within the class of ARIMA models. ARIMA models use historical or past values of a variable of interest, and/or the random error term as explanatory variables to forecast its future values. The variable of interest was a time series with equally spaced time intervals. Let's consider a discrete time series

$$Y_t = Y_1, Y_2 \quad (1)$$

The underlying principle of the Box-Jenkins ARIMA procedure was that it considers the observed time series Y_t as an output of inputs from an unobservable random process. These inputs are a series of independent random shocks e_t , which are assumed to be normally distributed with a zero mean and a constant variance, and referred to as white noise. In simple terms, the approach views a time series as a result of the transformation of a white noise process through a "linear filter" to obtain a particular set of outputs, which we now refer as the observed time series. ARIMA models of this form assume that the observed time series values may be dependent on;

The previous and current inputs (random shocks/white noise). The previous output values of the time series under study

$[Y_{t-1}, Y_{t-2}, \dots]$ in varying proportions

However, the Box-Jenkins approach assumes that the conditions at which the data was collected remain the same over time. If the assumption was not appropriate, a transfer function-noise model where a set of input variables which might have an effect on the time series are added to the model.

The Box-Jenkins ARIMA methodology is a five-step process for identifying, selecting, and assessing

Conditional mean models (for discrete, univariate time series data). The steps are listed below:

1. The researcher was establishing the stationarity of the time series. If the series is not stationary, researcher successively difference the series to attain stationarity. The sample autocorrelation function (ACF) and partial autocorrelation function (PACF) of stationary series decay exponentially (or cut off completely after a few lags).
2. The researcher was identifying a (stationary) conditional mean model for the data. The sample ACF and PACF functions can help with this selection. For an autoregressive (AR) process, the sample ACF decays gradually, but the sample PACF cuts off after a

few lags. Conversely, for a moving average (MA) process, the sample ACF cuts off after a few lags, but the sample PACF decays gradually. If both the ACF and PACF decay gradually, consider an ARMA model.

3. The researcher was specifying the model, and estimate the model parameters.
4. The researcher was conducting goodness-of-fit checks to ensure the model describes the data adequately. Residuals should be uncorrelated, homoscedastic, and normally distributed with constant mean and variance. If the residuals are not normally distributed, researcher can change the innovation distribution to a Student's t.
5. After choosing a model and checking its fit and forecasting ability, researcher used the model to forecast or generate Monte Carlo simulations over a future time horizon.

2.2. Model Building Strategy

Box-Jenkins ARIMA defined a four step iterative procedure for Model Identification, Model Estimation, Model Checking and Diagnosis, and Model Forecasting. The steps are described below.

Step 1: Model Identification

This step involves the identification of a tentative model, whether multiplicative or additive, and establishing the number of parameters involved and their combinations. This was done through analysis of historical data.

Visual inspection of time series plots was the first assessment tool. The first step was to consider the ACF and the PACF graphs to determine whether the series is stationary or not. Additionally, unit root tests are performed on the data to confirm stationarity and/or to make sure that differencing was necessary. If the series were non-stationary the series has to be either differenced to make it stationary in mean, or transformed if the covariance between any two observations Y_t and Y_{t+i} is not constant over time. However, differencing should be done with care to avoid the issue of over-differencing which might introduce dependence where none exists.

Secondly, a proposed model was estimated by finding the initial values (p , q , d) of the model parameters. This was done through looking at the significant coefficients in the ACF and PACF plots. The AC's and PAC's are compared with theoretical values to investigate candidate models. This procedure leads to deployment of various diagnostic tests which are conducted to first confirm stationarity of the series and check model fit. If the residual analysis confirms inadequacy of any particular model, a new model was proposed. The process was repeated until potential models are identified.

When the time series data has the seasonality component, seasonal differencing was recommended to make the data stationary. A seasonal difference was the difference between an observation Y_t and the corresponding observation from the previous year Y_{t-s} ; where s is the length of the season. It was recommended that the seasonal differencing be done before the first difference of the whole series as this might make the data stationary.

Step 2: Model Estimation

This is the process of estimating the model parameters after selecting a tentative model. The parameter estimates should be significant, with each providing a substantial contribution to the model for the most accurate forecasts. The researcher used a number of way to estimate autoregressive and moving averages parameters in ARMA models such as:

Maximum Likelihood Estimation Method

Maximum Likelihood Estimation

Given a time series Y_1, Y_2, \dots, Y_n the likelihood function L is defined as the joint probability of obtaining the observed series data. It was considered a function of the unknown model parameters with the observed data held fixed. Maximum likelihood estimators are those parameter values for which the actual data observed are most likely, that is, values that maximize the likelihood function.

Following parameter estimation, the most parsimonious model were selected using appropriate criterion from the pool of potential models identified. The Akaike Information Criterion (AIC) and the Schwarz Bayesian Information Criterion (BIC) are the two common goodness-of-fit statistics that are often used for model selection. The model with the lowest AIC and BIC is usually selected as the best fit.

The formula for the AIC is:

$$AIC = -2 \ln L + 2m \quad (2)$$

The BIC is given as:

$$BIC = -2 \ln L + (n) \quad (3)$$

Where: m is the number of model parameters ($m = p + q$);

n the number of observations, or equivalently, the sample size

Step 3: Model Diagnostics Checking

This step involves checking model adequacy, and if necessary incorporating potential improvements. Model checking was done through residual analysis. If the identified model is adequate, the residual observations should be transformed to a white noise process where the residuals are random and have the normal distribution. By studying the ACF plots of the residuals, the researcher established whether the AC's and PAC's are small and significant enough to consider the model adequate. If the autocorrelations are large, the values of p and/or q are adjusted and the model re-estimated until the best fit model is estimated.

The residuals of an ARMA (p , q) can be obtained as below:

$$\hat{e}_t = \hat{Y}_t - \left(\hat{\delta} + \sum_{i=1}^p \hat{\theta}_i Y_{t-i} - \sum_{i=1}^q \hat{\theta}_i \hat{e}_{t-i} \right) \quad (4)$$

Step 4: Forecasting

Forecasting was the last stage after the model has been identified and fitted. The researcher used the model to generate forecasts of future values. If we denote the current time as t , the forecast for \hat{Y}_{t+k} is the k period-ahead forecast

denoted by $\hat{Y}_{t+k}(t)$.

For an ARIMA (p, d, q) process at time $t + k$ (k periods in the future) the model is:

$$Y_{t+k} = \delta + \sum_{i=1}^{p+d} \phi_i Y_{t+k-i} + \epsilon_{t+k} - \sum_{i=1}^q \theta_i \epsilon_{t+k-i} \quad (5)$$

The variance of the forecast error gets bigger with increasing forecast lead times k .

2.3. Evaluating Model Performance

Once the model has been fitted, subjected to diagnostic checks and used to produce forecasts, it was evaluated with forecast fit measures. This study uses the maximum absolute percentage error (MaxAPE) and the maximum absolute error

(MaxAE) to evaluate the amount of this forecast error. These measures were computed using the proposed model both for the original and forecasted series. The model with least forecast errors was considered accurate.

3. Results

Building ARIMA model for cancer case incidents data and Forecasting:

To fit an ARIMA model requires a sufficiently large data set. In this study, the data for cancer case incidents for the period 2000-2015 were used. As earlier stated that development of ARIMA model for any variable involves three steps: identification, estimation and verification. Each of these three steps is now explained for cancer case incidents.

Table 1. Cancer cases incidents in different hospitals.

Year	Narok Hospital	Longisa Hospital	Tenwek Hospital	Kapkatet Hospital	Litein Hospital	Cancer cases (Total in number)
2000-01	4	15	20	12	6	57
2001-02	5	14	23	10	10	61
2002-03	3	12	18	7	11	51
2003-04	6	8	11	9	10	44
2004-05	8	14	16	6	14	58
2005-06	7	5	22	20	6	60
2006-07	9	14	23	10	13	69
2007-08	12	15	21	12	11	71
2008-09	11	18	20	10	14	73
2009-10	13	22	19	13	10	77
2010-11	17	27	30	10	20	110
2011-12	15	24	25	20	19	103
2012-13	13	22	24	11	21	91
2013-14	16	20	24	19	20	104
2014-15	19	26	30	25	21	121

3.1. Model Identification

ARIMA model is estimated only after transforming the variable under forecasting into a stationary series. The stationary series is the one whose values vary over time only around a constant mean and constant variance. There are several ways to ascertain this. The most common method is to check stationarity through examining the graph or time plot of the data. Fig1 reveals that the data is nonstationary. Non-stationarity in mean is corrected through appropriate differencing of the data. In this case difference of order 1 was sufficient to achieve stationarity in mean.

The newly constructed variable X_t can now be examined for stationarity. The graph of X_t was stationary in mean. The next step is to identify the values of p and q . For this, the autocorrelation and partial autocorrelation coefficients of various orders of X_t are computed (Table 2). The ACF and PACF (figure. 2 and 3) shows that the order of p and q can at most be 1. We entertained three tentative ARIMA models and chose that model which has minimum AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion).

The models and corresponding AIC and BIC values are

ARIMA (p, d, q)	AIC	BIC
210	409.759	415.612
211	410.289	418.094
212	412.358	422.114

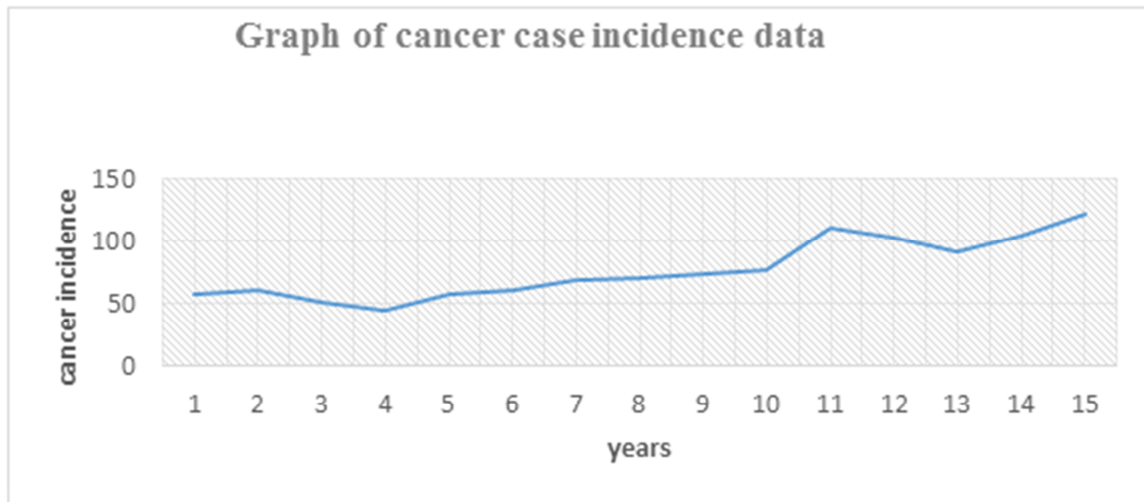
So the most suitable model is ARIMA (2, 1, 0) as this model has the lowest AIC and BIC values.

3.2. Model Estimation and Diagnostic Checking

Model parameters were estimated using SPSS package. Results of estimation are reported in table 3. The model verification is concerned with checking the residuals of the model to see if they contain any systematic pattern which still can be removed to improve on the chosen ARIMA. This is done through examining the autocorrelations and partial autocorrelations of the residuals of various orders. For this purpose, the various correlations up to 12 lags were computed and the same along with their significance which is tested by Box-Ljung test are provided in table 4. As the results indicate, none of these correlations is significantly different from zero at a reasonable level. This proves that the selected ARIMA model is an appropriate model.

The ACF and PACF of the residuals (figure. 4 and 5) also

indicate 'good fit' of the model. So the fitted ARIMA model $Y_t = 4.6022 + 1.1209 Y_{t-1} - .7630 Y_{t-2} + .6421 Y_{t-3} + \varepsilon_t$ (6) for the cancer data is



YEAR, Not Periodic

Figure 1. Time plot of cancer incidence case data.

ACF of differenced data

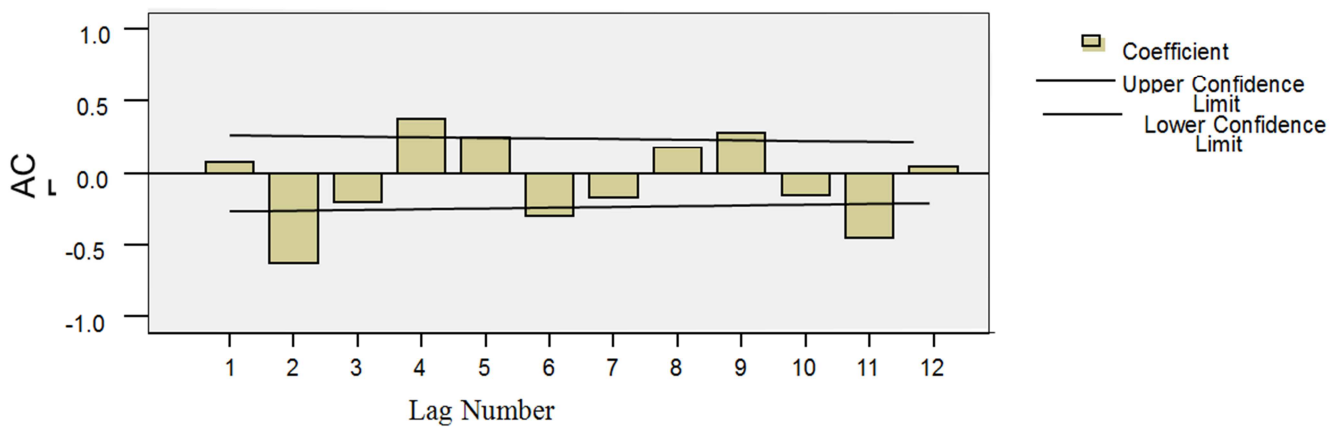


Figure 2. ACF of differenced data.

PACF of differenced data

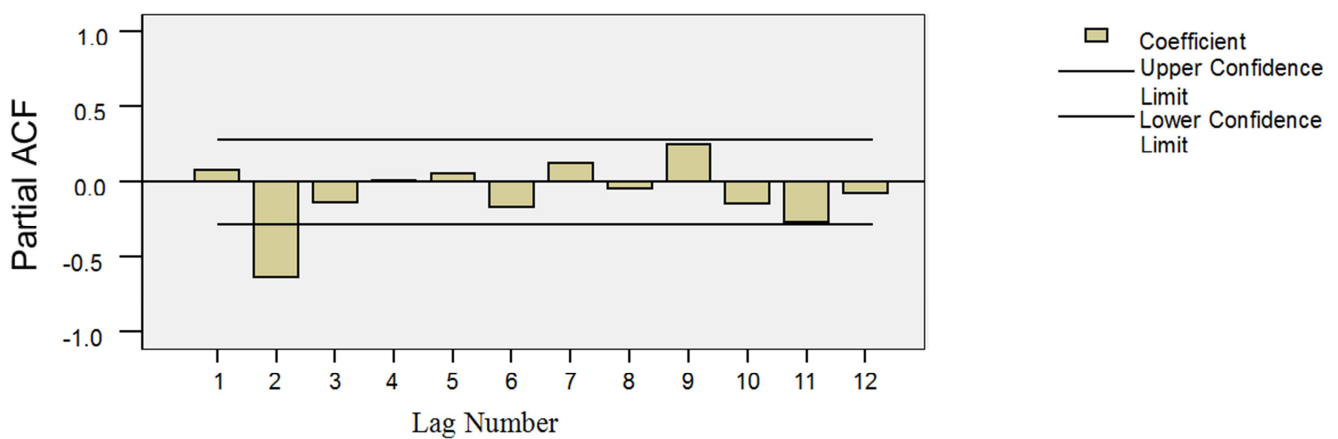


Figure 3. PACF of differenced Cancer case incidence data.

Table 2. Autocorrelations and partial autocorrelations.

Lag	Autocorrelation	Std.Error	Lag	Partial Autocorrelation	Std.Error
1	0.074	0.135	1	0.074	0.139
2	-0.623	0.133	2	-0.632	0.139
3	-0.203	0.132	3	-0.137	0.139
4	0.373	0.131	4	0.010	0.139
5	0.252	0.129	5	0.049	0.139
6	-0.295	0.128	6	-0.169	0.139
7	-0.172	0.127	7	0.117	0.139
8	0.178	0.125	8	-0.043	0.139
9	0.276	0.124	9	0.249	0.139
10	-0.153	0.122	10	-0.143	0.139
11	-0.454	0.121	11	-0.268	0.139
12	0.045	0.119	12	-0.077	0.139

Table 3. Estimates of the fitted ARIMA model.

		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	AR1	0.1209	0.1093	1.1055	0.2743
	AR2	-0.6421	0.1087	-5.9069	0.0000
Constant		4.6022	1.1076	4.1551	0.0001
Number of Residuals	15				
Number of Parameters	2				
Residual df	13				
Adjusted Residual Sum of Squares	7166.782				
Residual Sum of Squares	7168.028				
Residual Variance	143.286				
Model Std. Error	11.97021				
Log-Likelihood Akaike's Information Criterion (AIC)	-201.879				
Schwarz's Bayesian Criterion (BIC)	409.7587				
	415.6125				

Table 4. Autocorrelation and partial autocorrelations of residuals.

Lag	Autocorrelation	Std.Error	Box-Ljung Statistic	df	Sig.
			Value		
1	-0.0898	0.1348	0.4438	1.0000	0.5053
2	-0.0156	0.1334	0.4575	2.0000	0.7955
3	-0.0524	0.1321	0.6149	3.0000	0.8930
4	-0.1268	0.1307	1.5559	4.0000	0.8167
5	0.1815	0.1294	3.5250	5.0000	0.6196
6	-0.1441	0.1280	4.7930	6.0000	0.5706
7	0.1967	0.1266	7.2080	7.0000	0.4075
8	-0.0612	0.1252	7.4474	8.0000	0.4892
9	-0.0335	0.1237	7.5208	9.0000	0.5831
10	-0.0292	0.1223	7.5780	10.0000	0.6700
11	-0.2872	0.1208	13.2276	11.0000	0.2787
12	0.0616	0.1194	13.4940	12.0000	0.3342

Partial		
Lag	Autocorrelation	Std. Error
1	-0.0898	0.1387
2	-0.0239	0.1387
3	-0.0565	0.1387
4	-0.1389	0.1387
5	0.1583	0.1387
6	-0.1309	0.1387
7	0.1831	0.1387
8	-0.0518	0.1387
9	0.0062	0.1387
10	-0.0870	0.1387
11	-0.2323	0.1387
12	-0.0695	0.1387

ACF of residuals

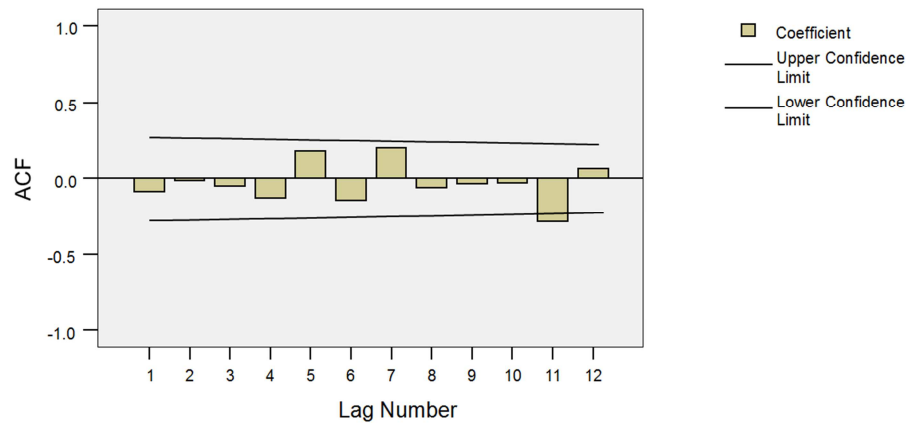


Figure 4. ACF of residuals of fitted ARIMA model.

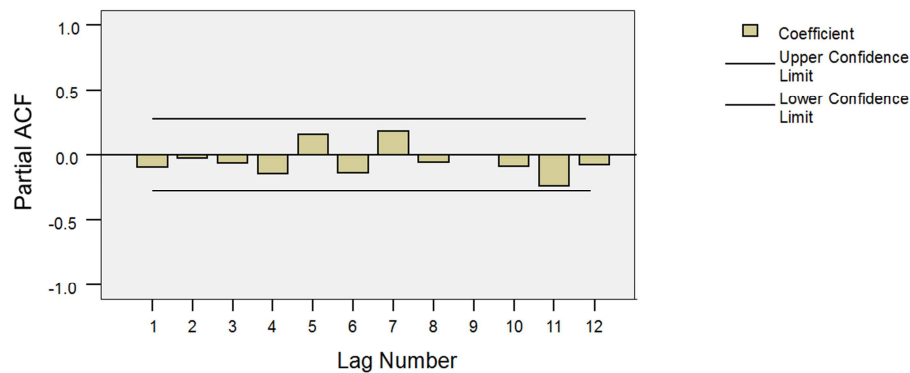


Figure 5. PACF of residuals of fitted ARIMA model.

3.3. Forecasting with ARIMA Model

ARIMA models are developed basically to forecast the corresponding variable. There are two kinds of forecasts: sample period forecasts and post-sample period forecasts. The former are used to develop confidence in the model and the latter to generate genuine forecasts for use in planning and other purposes. The ARIMA model can be used to yield both these kinds of forecasts.

3.3.1. Sample Period Forecasts

The sample period forecasts are obtained simply by plugging the actual values of the explanatory variables in the estimated equation (1.5). The explanatory variables here are the lagged values of Y_t and the estimated lagged errors. Then so obtained values for Y_t together with the actual values of Y_t are shown in table 5.

Table 5. Actual and estimated values of cancer case incidents and 95% confidence limit (CL).

Year	Actual cancer case incidence	Estimated cancer case incidence	Residual	Lower CL	Upper CL
2000-01	57
2001-02	61	61.6	-0.022	30.110	93.195
2002-03	51	66.231	-15.231	34.784	97.677
2003-04	44	53.775	-9.365	29.483	78.067
2004-05	58	57.440	1.300	33.148	81.732
2005-06	60	71.705	-11.165	47.413	95.997
2006-07	69	58.557	10.493	34.265	82.849
2007-08	71	75.924	-4.764	51.632	100.216
2008-09	73	72.952	0.408	48.660	97.244
2009-10	77	79.272	-1.452	54.980	103.564
2010-11	110	83.948	26.052	59.655	108.240
2011-12	103	118.027	-14.057	93.735	142.319
2012-13	91	89.579	2.331	65.287	113.871
2013-14	104	101.325	2.905	77.033	125.617
2014-15	121	120.464	1.446	96.172	144.756
2015-16		123.137		98.845	147.429
2016-17		125.890		100.598	150.182
2017-18		126.729		102.437	153.021

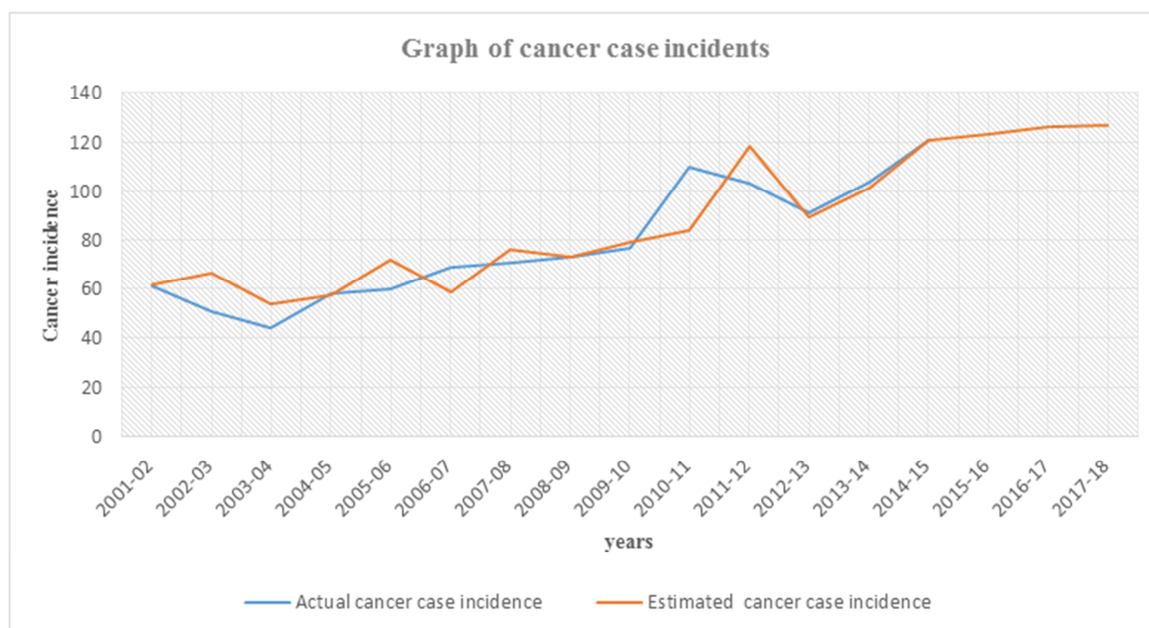


Figure 6. Actual and estimated cancer case incidents.

3.3.2. Post Sample Forecasts

The principal objective of developing an ARIMA model for a variable is to generate post sample period forecasts for that variable. This was done through using equation (1.5). The forecasts for cancer case incidents during 2016 to 2018 are given in lower part of table 5.

4. Conclusion

To judge the forecasting ability of the fitted ARIMA model, important measures of the sample period forecasts' accuracy were computed. The Mean Absolute Percentage Error (MAPE) for cancer case incidents turns out to be 6.264337. This measure indicates that the forecasting inaccuracy is low.

ARIMA model offers a good technique for predicting the magnitude of any variable. Its strength lies in the fact that the method is suitable for any time series with any pattern of change and it does not require the forecaster to choose a priori the value of any parameter. Its limitations include its requirement of a long time series. Often it is called a 'Black Box' model. Like any other method, this technique also does not guarantee perfect forecasts. Nevertheless, it can be successfully used for forecasting long time series data.

5. Discussion

In this study the developed model for cancer case incidents was found to be ARIMA (2, 1, 0). From the forecast available by using the developed model, it can be seen that forecasted incidents for the year 2015-16 is higher than 2014-15 and in later years the incidents increases. The validity of the forecasted values can be checked when the data for the lead periods become available. The model can be used by

researchers for forecasting of cancer incidents in Kenya. However, it should be updated from time to time with incorporation of current data.

Dedication

This research paper is dedicated to my loving wife Rose Langat and my son Ramsey Kipkirui

Acknowledgement

I acknowledge the support of Prof. J. K Sigei and Prof. Romanus Odhiambo whose technical support has seen me through in the entire period of this research work.

And above all God, the creator, giver of all that is good for seeing me through this process.

References

- [1] American Cancer Society: Cancer in Africa. Atlanta, GA, American Cancer Society, 2011.
- [2] Danaei G et al. 2005, Causes of cancer in the world: Comparative assessment of nine behavioural and environmental risk factors, *Lancet* 330: 223.
- [3] Stefan DC, Elzawawy AM, Khaled HM, et al: Developing cancer control plans in Africa: Examples from five countries. *Lancet Oncol* 14: e189-e195, 2013.
- [4] Doll R and Peto R. The causes of cancer. *Journal of the National Cancer Institute*, 1981, 66: 1191-308.
- [5] National Cancer Control Programmes: Policies and management guidelines. Geneva, World Health Organization, 1995. Sobin et al. TNM classification of malignant tumours, 7th edition 2009 John Wiley and Sons.

- [6] Textbook of Therapeutics: Drug and Disease Management Seventh Edition edited by Eric T. Herfindal and Dick R. Gourley.
- [7] Wiley Textbook of Uncommon Cancer, 2nd Edition D. Raghavan (Editor), M. L. Brecher (Editor), D. H. Johnson (Editor), N. J. Meropol (Editor), P. L. Moots (Editor), J. T. Thigpen (Editor) ISBN: 0-471-92921-2, 776 pages.
- [8] KNBS: Kenya Facts and Figures 2012. Kenya National Bureau of Statistics, 2014.
- [9] KNBS: Kenya Population Census (2009). Preliminary Report. Kenya National Bureau of Statistics, 2009.
- [10] Kenya National Cancer Control strategy 2011-16.
- [11] Driscoll T et al 2005. The global burden of diseases due to occupational carcinogens. American journal of Industrial Medicine, 48:419-431.
- [12] Jayant K et al 1998, Survival from cervical cancer in Barchi registry, rural India. Cancer Survival in developing countries IARC Scientific Publication No 145 pp 69-77.
- [13] Matthers CD, Loncar D 2006, projections of global mortality and burden of disease from 2002 to 2030. PloS Medicine, 3: 2011-2030.
- [14] Ponten J et al. 1995, Strategies for control of cervical cancer. International journal of cancer, 60 1-26.
- [15] KHPF: Analysis of Performance, Analytical Review of Health Progress, and Systems Performance 1994-2010, Kenya. 2010.
- [16] Van Hemelrijck MJ, Lindblade KA, Kubaje A, Hamel MJ, Odhiambo F, Phillips-Howard PA, et al: Trends observed during a decade of paediatric sick visits to peripheral health facilities in rural western Kenya, 1997-2006. Trop Med Int Health 2009, 14: 62-69.
- [17] Chatfield C: The analysis of time series: an introduction. London: Chapman & Hall; 2004.
- [18] Montgomery DC, Jennings, C. L, Kulahci, M.: Introduction to Time Series and Forecasting. Wiley & Sons; 2008.
- [19] Makridakis S, Wheelwright, SC. & Hyndman, RJ.: Forecasting Methods and Applications. New York: John Wiley & Sons, Inc; 1998.
- [20] Shumway RH, S, D. S.: Time Series Analysis and Its Applications. Springer; 2009.
- [21] Abeku TA, de Vlas SJ, Borsboom G, Teklehaimanot A, Kebede A, Olana D, et al: Forecasting malaria incidence from historical morbidity patterns in epidemic-prone areas of Ethiopia: a simple seasonal adjustment method performs best. Trop Med Int Health 2002, 7:851-857.
- [22] Briet OJ, Vounatsou P, Gunawardena DM, Galappaththy GN, Amerasinghe PH: Models for short term malaria prediction in Sri Lanka. Malar J 2008, 7: 76.
- [23] Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White NJ, Kaewkungwal J: Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: a case study in endemic districts of Bhutan. Malar J 2010, 9: 251.
- [24] Kumar V, Mangal, A, Panestar, S., et al Forecasting Malaria Cases Using Climatic Factors in Delhi, India: A Time Series Analysis. Malaria Research and Treatment 2014, 2014.
- [25] Ezekiel D, Opara, J. & Idochi, O: Modelling and Forecasting Malaria Mortality Rate using SARIMA Models (A Case Study of Aboh Mbaise General Hospital, Imo State Nigeria). Science Journal of Applied Mathematics and Statistics 2014, 2: 31-41.
- [26] Zhang X, Zhang, T., Young, AA. & Li, X.: Applications and Comparisons of Four Time Series Models in Epidemiological Surveillance Data. PLoS ONE 2014, 9.
- [27] Briet OJ, Vounatsou P, Gunawardena DM, Galappaththy GN, Amerasinghe PH: Models for short term malaria prediction in Sri Lanka. Malar J 2008, 7: 76.
- [28] Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White NJ, Kaewkungwal J: Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: a case study in endemic districts of Bhutan. Malar J 2010, 9: 251.
- [29] Zhang X, Zhang, T., Young, AA. & Li, X.: Applications and Comparisons of Four Time Series Models in Epidemiological Surveillance Data. PLoS ONE 2014, 9.
- [30] Caldwell, J. G. (n.d.) The Box-Jenkins forecasting technique. Retrieved March 3, 2012, from <http://www.foundationwebsite.org/BoxJenkins.htm>.
- [31] Rycroft, R. S. (1995). Student editions of forecasting software: A survey. International Journal of Forecasting, 11, 337-351.
- [32] Rowe AK: Potential of integrated continuous surveys and quality management to support monitoring, evaluation, and the scale-up of health interventions in developing countries. Am J Trop Med Hyg 2009, 80: 971-979.
- [33] The GLOBOCAN Project, estimated cancer incidence, prevalence, mortality and DALYs Worldwide in 2008, World Health Organization.
- [34] Allard R: Use of time-series analysis in infectious disease surveillance. Bull World Health Organ 1998, 76: 327-333.
- [35] Geurts, M. D., & Ibrahim, I. B. (1975). Comparing the Box-Jenkins approach with the exponentially smoothed forecasting model application to Hawaii tourists. Journal of Marketing Research, 12 (2), 52-66.
- [36] Levenback, H., & Cleary, J. P. (2006). Forecasting practice and process for demand management. Belmont, CA: Thomson Brooks/Cole.
- [37] Box, G. E., & Jenkins, G. M. (1994). Time series analysis: Forecasting and control (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
- [38] Uziel, A., & Stone, L. (2012). Determinants of periodicity in seasonally driven epidemics. Journal of Theoretical Biology, 305, 88-95.
- [39] GoK.: Kenya Population Situation Analysis. Kenya: Government of Kenya 2013.
- [40] Pankratz, A. (1983). Forecasting with univariate Box-Jenkins concepts and cases. New York, NY: John Wiley & Sons.
- [41] Makridakis, S., & Wheelwright, S. C. (Eds.). (1987). The handbook of forecasting. A manager's guide (2nd ed.). New York, NY: John Wiley & Sons.

- [42] Chatfield C: The analysis of time series: an introduction. London: Chapman & Hall; 2004.
- [43] Dickey DF, WA: Distribution of the Estimators for Autoregressive Time Series With a Unit Root. Journal of the American Statistical Association 1979, 74: 427-431.
- [44] Box GEPJ, G. M: Time Series Analysis: Forecasting and Control. San Fransisco, Holden day 1976.
- [45] Slutsky's E: The summation of random causes as the source of cyclic processes. Econometrica 1937, 5: 105-146.
- [46] Wold H: A Study in the Analysis of Stationary Time Series. 2nd. Ed. edn. Stockholm: Almqrist & Wiksell; 1954.
- [47] Pankratz A: Forecasting with Univariate Box–Jenkins Models: Concepts and Cases. New York: John Wiley & Sons; 1983.
- [48] Ljung GB, GEP.: On a measure of lack of fit in time series models. Biometrika 1978, 2: 297-303.
- [49] Hyndman RK, AB.: Another look at measures of forecast accuracy. Int J Forecasting 2006, 22: 679– 688.
- [50] Armstrong JS CF: Error measures for generalizing about forecasting methods—empirical comparisons. Int J Forecasting 1992, 8: 69-80.