academicJournals

Vol. 8(6), pp. 136-142, June, 2013
DOI: 10.5897/AJMCSR2013.0462
ISSN 2006-9731 © 2013 Academic Journals
http://www.academicjournals.org/AJMCSR

African Journal of Mathematics and Computer Science Research

Full Length Research Paper

Incidents of cancer in Sudan: Past trends and future forecasts

Ehab Ahmed Mohammed[1*], Atif Alagib[1], and Arbab Ismail Babiker[2]

[1]Tropical Medicine Research Institute, National Center for Research.
[2]Department of Econometric and Social Statistics, Faculty of Economic and Social Studies, University of Khartoum, Khartoum, Sudan.

Accepted 17 May, 2013

Incidents of cancer in Sudan have been growing in numbers over the last five decades (1967-2010). Using data compiled regularly by radiations isotope of cancer in Khartoum (RICK) - which continued to be the only cancer treatment centre in Sudan for over the last half century- its trend is studied using Box-Jenkins methodology in time series analysis is the optimal method applied to the pattern. This method consists of four steps namely identification, estimation, diagnostic checking, and forecasting by autoregressive integrated moving average (ARIMA) models. Future forecasts drawn show that the number of incidents is likely to continue growing if no significant intervention is made by the health authorities as intervention measures are undertaken to curb it.

Key words: Radiations isotope of cancer in Khartoum (RICK), cancer, autoregressive integrated moving average (ARIMA) model, stationarity.

INTRODUCTION

Despite the fact that data on health are generally very scanty, fairly good data on cancer incidents have been compiled via continuous registration systems at hospital levels. Good sources of data on cancer incidents over the years have been those records of patients attending the only ontological hospital offering specialized treatment for cancer in the country, the radiation and isotopes centre, Khartoum (RICK), and the Sudan cancer registry based on histopathologically confirmed cases diagnosed in the National Health Laboratories in Khartoum (Parkin et al., 2003). Early reports presented data on histopathologically confirmed cases. For example, Hickey (1959) reported 1,335 malignant epithelial neoplasms collected from the Stack Medical Research Laboratories during the period 1935-1954, El Hassan and Lynch (1962) reported 2,234 malignant tumors collected from the same source and from the Department of Pathology, University of Khartoum, for the period 1954-1961. This series was reproduced by Daoud et al. (1968) and compared with 1,578 malignant tumors from Khartoum district examined at the Department of Pathology in 1957-1965.

Table 1 and its corresponding pictorial representation Figure 1, shows the general growth pattern of incidents of cancer in Sudan as compiled from different data sources which is approximately an exponential trend.

A cursory look at these data coupled with some calculations reveal that cancer incidents in Sudan have been growing at an average annual rate of 0.061 over the period covered. Taking source of data into consideration, the considerable growth in the number of incidents of cancer in Sudan over the last decade or so, as observed from Figure 1, may be attributed to increasing awareness

*Corresponding author. E-mail: ehabfrah@hotmail.com.

Table (1). Incidents of Cancer in Sudan (1967- 2010).

Year	Number of incidents of cancer	Cancer rate/1000	Year	Number of Incidents of cancer	Cancer rate/1000
1967	303	0.0234	1989	1357	0.0558
1968	448	0.0339	1990	1572	0.0629
1969	540	0.0400	1919	1494	0.0582
1970	512	0.0371	1992	2157	0.0817
1971	538	0.0382	1993	1847	0.0722
1972	500	0.0348	1994	1645	0.0625
1973	562	0.0398	1995	1733	0.0640
1974	692	0.0472	1996	1810	0.0649
1975	470	0.0308	1997	2119	0.0739
1976	565	0.0357	1998	2145	0.0727
1977	738	0.0449	1999	2102	0.0692
1978	545	0.0319	2000	2541	0.0813
1979	568	0.0320	2001	2963	0.0922
1980	704	0.0381	2002	3070	0.0928
1981	672	0.0350	2003	3185	0.0936
1982	773	0.0388	2004	3450	0.0986
1983	870	0.0422	2005	3705	0.1029
1984	913	0.0431	2006	3505	0.0946
1985	903	0.0415	2007	4813	0.1262
1986	1112	0.0497	2008	5156	0.1317
1987	927	0.0403	2009	5739	0.1425
1988	1308	0.0553	2010	6303	0.1522

among patients of the need to visit specialized hospitals for diagnosis and treatment, where they will be registered via the continuous registration system. This pointed to the seriousness of the situation and necessitated the investigation of volume of future incidents which will occur if similar conditions prevail. These number incidents rate expressed as rates per 1000 population are also shown in Table 1. Its annual rate of growth over the period covered is 0.03443 which is fairly similar to its counterparts for other common diseases in Sudan such as malaria (FMOH, 2010).

Cancer incidents model

As is generally known, developing a time series model from such data starts by exploring the main features inherent in the series. Among these features are stationarity and the existence of seasonality (cyclical pattern) in the data. Appropriate statistical procedures will now be used for investigating these aspects of the series in an attempt to determine the suitable time series model that fits it.

Testing for stationarity

Stationary series vary around the constant mean level, neither decreasing nor increasing systematically over time with constant variance. Certain time series models, like the Box-Jenkins model, assume the existence of stationarity. General Box-Jenkins model includes difference operators, autoregressive terms, moving average terms, seasonal difference operators, seasonal autoregressive terms, and seasonal moving average terms. This phase is founded on the study of autocorrelation and partial autocorrelation.

The Box-Jenkins model assumes the stationarity of the series under investigation, which means that the series has constant mean, constant variance, and constant autocorrelation structure. Thus, the first step in developing a Box-Jenkins model is to determine if the series is stationary and if there is any significant seasonality that needs to be modeled (Box and Jenkins, 1970).

Consider the AR (1) model:

$$y_t - \mu = \phi_1 (y_{t-1} - \mu) + a_t$$

For this model the autoregressive polynomial equation is $1 - \phi_1 z = 0$ and therefore is the root of the autoregressive polynomial. Thus, for the AR (1) model to be stationarity, it is required that $z_1^{AR} = \frac{1}{\emptyset_1 \left|\frac{1}{\emptyset_1}\right|} > 1$ and therefore $|\emptyset_1| < 1$.

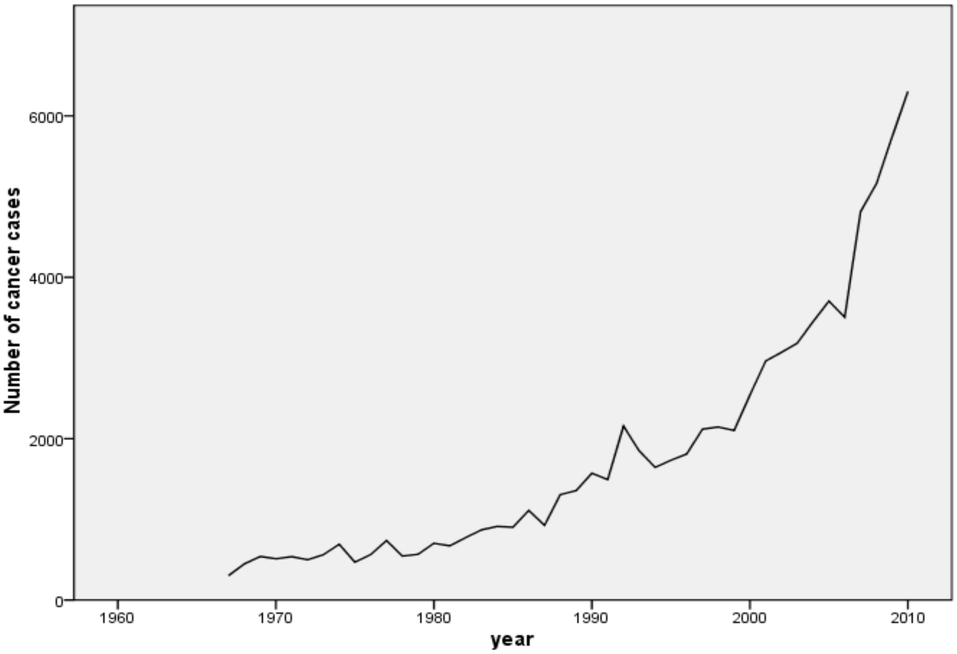

Figure 1. Incidents of cancer in Sudan (1967-2010).

Similarly, for an MA (1) model $y_t - \mu = a_t - \theta_1 a_{t-1}$ to be invertible, it is required that $z_1^{AR}|1/\theta_1|$ and therefore $|\theta_1| < 1$.

For the stationarity and invertibility conditions for other popular Box-Jenkins models like the AR (2), MA (2), and ARMA (1, 1) models, see ADF and PACF results.

By definition, all AR (p) models are invertible while all MA (q) models are stationarity. Now consider the practical implications of stationarity and invertibility in Box-Jenkins models. When a Box-Jenkins model is stationarity, its observations y_t satisfies the following three properties:

(1) $E(y)_t$ =μ (that is, the mean of $\forall_t y_t$ is constant for all time periods)
(2) Var(y_t) = σ_y^2 (that is, the variance of $\forall_t y_t$ is constant for all time periods)
(3) Cov($y_t y_{t-j}$) = γ_j (that is, the covariance between y_t and y_{t-j} is constant for all time periods and fixed j, j = 1, 2,)

These three conditions give rise to what is called weak stationarity (or just stationarity for short). The practical implication of stationarity is that only one realization of the time series y_t is needed for us to be able to consistently estimate the mean μ, the variance σ_y^2, the covariance γ_j, and the autocorrelation ρ_j with the sample statistics $\bar{y}, s_y^2, c_j$ and r_j. These statistics are defined as:

$$\bar{y} = \frac{\sum_{t=1}^{T} y_t}{T} \quad r_j \tag{1}$$

$$S_y^2 = \frac{\sum_{t=1}^{T} (y_t - \bar{y})^2}{T} \tag{2}$$

$$C_j = \frac{\sum_{t=1}^{T} (y_t - \bar{y})(y_{t-j} - \bar{y})}{T} \tag{3}$$

$$r_j = \frac{c_j}{S_y^2} = \frac{\sum_{t=j+1}^{T} (y_t - \bar{y})(y_{t-j} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})} \tag{4}$$

where *T* denotes the total number of observations available on y_t (sample mean) for Equation 1, Equations 2 is the Sample variance; 3 is Sample covariance and 4 Sample autocorrelation. Stationarity can be accessed from a run sequence plot. The run sequence plot should show constant location and scale. It can also be detected from an autocorrelation plot. Specifically, non-stationarity is often indicated by an autocorrelation plot with very slow decay.

Box and Jenkins recommend differencing non-stationary series one or more times to achieve stationarity. Doing so produces an Autoregressive integrated moving average (ARIMA) model, with "I" short for "Integrated". But its first difference, expressed as $\Delta \mathrm{y}_{\mathrm{t}} = y_{\mathrm{t}} - y_{t-1} = u_t$, is stationary, so y is integrated of order 1", or $y \sim \mathrm{I}$.

Seasonality in Box-Jenkins models

Box-Jenkins models can be extended to include seasonal autoregressive and seasonal moving average terms.

Model identification: seasonality of order s is revealed by "spikes" at s, 2s, 3s, lags of the autocorrelation function.

Model estimation: to make a series stationary, may need to take sth differences of the raw data before estimation. These seasonal effects may themselves follow AR and MA processes.

At the model identification stage, our goal is to detect seasonality, if it exists, and to identify the order for the seasonal autoregressive and seasonal moving average terms. For Box-Jenkins models, it isn't necessary to remove seasonality before fitting the model. Instead, it can include the order of the seasonal terms in the model specification to the ARIMA estimation software.

Once stationarity and seasonality have been addressed, the next step is to identify the order (the p and q) of the autoregressive and moving average terms. The primary tools for doing this are the autocorrelation plot and the partial autocorrelation plot. The sample autocorrelation plot and the sample partial autocorrelation plot are compared to the theoretical behaviour of these plots when the order is known.

Order of autoregressive process (p)

Specifically, for an AR (1) process, the sample autocorrelation function should have an exponentially decreasing appearance. However, higher-order AR processes are often a mixture of exponentially decreasing and damped sinusoidal components. For higher-order autoregressive processes, the sample autocorrelation needs to be supplemented with a partial autocorrelation plot. The partial autocorrelation of an AR (p) process becomes zero at lag p+1 and greater, so we examine the sample partial autocorrelation function to see if there is evidence of a departure from zero. This is usually determined by placing a 95% confidence interval on the sample partial autocorrelation plot (most software programs that generate sample autocorrelation plots will also plot this confidence interval). If the software program does not generate the confidence band, it is approximately ±2/N, with N denoting the sample size.

The data is AR (p) if: autocorrelation function (ACF) will decline steadily, or follow a damped cycle and partial autocorrelation function (PACF) will cut off suddenly after p lags.

Order of moving average process (q)

The autocorrelation function of an MA (q) process becomes zero at lag q+1 and greater, so we examine the sample autocorrelation function to see where it essentially becomes zero. Alternating positive and negative, autoregressive model. Using the partial autocorrelation plot to decaying to zero, help identify the order as one or more spikes; the rest are Moving average model, where order is identified by where plot essentially zero, becomes zero. Decay, starting after a few lags mixed autoregressive and moving average model. All zero or close to zero data is essentially random. High values at fixed intervals Include seasonal autoregressive term. No decay to zero series is stationary.

The data is MA (q) if: ACF will cut off suddenly after q lags and PACF will decline steadily, or follow a damped cycle.

It is not indicated to build models with:

(1) Large numbers of MA terms

(2) Large numbers of AR and MA terms together; you may well see very (suspiciously) high t-statistics. This happens because of high correlation ("co linearity") among regressors, not because the model is good.

It is observable from Figure 2 that the time series is likely to have random walk pattern. Moreover, ACFs suffered from linear decline and there is only one significant spike for PACFs. The correlogram also suggests that ARIMA (1, 0, 0) may be an appropriate model. Then, we take the first-difference of "cancer" to see whether the time series becomes stationary before further finding AR (p) and MA (q).

To see whether first difference can get level-stationary time series or not, the results are: the first-difference series "cancer" becomes stationary as shown in line graph (Figure 3) and is white noise as it shows no significant patterns in the graph of correlogram (Figure 4). And the unit root test also confirms the first-difference becomes stationary since the ADF value is less than 1% critical value l, (Tables 2 and 3).

Box-Jenkins model estimation

The main approaches to fitting Box-Jenkins models are non-linear least squares and maximum likelihood estimation. Maximum likelihood estimation is generally the preferred technique (Box et al., 1994).

Box-Jenkins model diagnostics

Model diagnostics for Box-Jenkins models is similar to

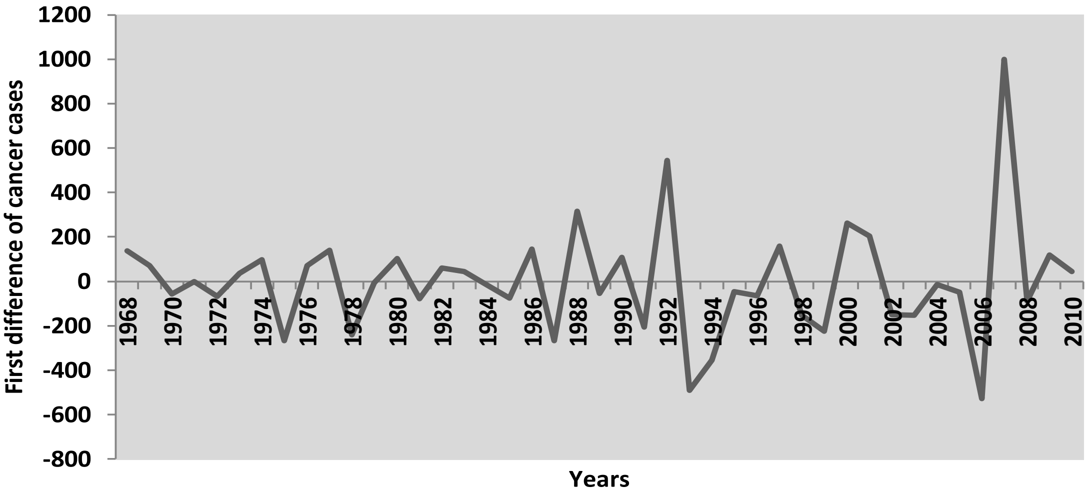

Figure 2. First difference trend of cancer series in Sudan (1967-2010).

Table 2. Correlogram graph of incidents of cancer in Sudan (1967-2010).

Auto correlation	Partial correlation		AC	PAC	Q-Stat	Prob
. |*******|	. |*******|	1	0.866	0.866	35.333	0.000
. |****** |	. | . |	2	0.754	0.012	62.708	0.000
. |***** |	. | . |	3	0.654	-0.004	83.855	0.000
. |**** |	. | . |	4	0.556	-0.048	99.517	0.000
. |**** |	. |*. |	5	0.509	0.146	112.98	0.000
. |*** |	.*| . |	6	0.447	-0.070	123.63	0.000
. |*** |	. | . |	7	0.388	-0.021	131.84	0.000
. |*** |	. | . |	8	0.334	-0.018	138.13	0.000
. |** |	. | . |	9	0.276	-0.024	142.53	0.000
. |** |	.*| . |	10	0.214	-0.074	145.27	0.000
. |*. |	. | . |	11	0.168	0.018	147.00	0.000
. |*. |	. | . |	12	0.137	0.030	148.19	0.000
. |*. |	.*| . |	13	0.095	-0.068	148.79	0.000
. | . |	. | . |	14	0.054	-0.045	148.99	0.000
. | . |	. | . |	15	0.032	0.056	149.06	0.000
. | . |	. | . |	16	0.003	-0.033	149.06	0.000
. | . |	. | . |	17	-0.020	-0.028	149.09	0.000
. | . |	.*| . |	18	-0.056	-0.078	149.33	0.000
.*| . |	.*| . |	19	-0.112	-0.100	150.35	0.000
.*| . |	. | . |	20	-0.143	0.018	152.08	0.000

model validation for non-linear least squares fitting. That is, the error term u_t is assumed to follow the assumptions for a stationary unvaried process. The residuals should be white noise (or independent when their distributions are normal) drawings from a fixed distribution with a constant mean and variance.

If the Box-Jenkins model is a good model for the data, the residuals should satisfy these assumptions. If these assumptions are not satisfied, we need to fit a more appropriate model. That is, we go back to the model identification step and try to develop a better model. Hopefully the analysis of the residuals can provide some clues as to a more appropriate model. The residual analysis is based on:

$$\mathbf{Q(S)} = n \sum r(k)^2 \approx \chi^2 \ \mathbf{(S)} \qquad (5)$$

(1) Random residuals: the Box-Pierce Q-statistic: where

Table 3. Correlogram graph of first difference cancer series.

Auto correlation	Partial correlation		AC	PAC	Q-Stat	Prob
. | . |	. | . |	1	-0.027	-0.027	0.0338	0.854
. |** |	. |** |	2	0.213	0.212	2.1758	0.337
. |** |	. |** |	3	0.215	0.237	4.4206	0.219
. |*. |	. |*. |	4	0.096	0.079	4.8741	0.300
. | . |	.*| . |	5	0.000	-0.094	4.8741	0.431
. |*. |	. | . |	6	0.138	0.049	5.8753	0.437
. |*. |	. |*. |	7	0.074	0.078	6.1681	0.520
. | . |	. | . |	8	0.057	0.047	6.3470	0.608
. |*. |	. |*. |	9	0.127	0.076	7.2663	0.609
. | . |	. | . |	10	0.047	-0.005	7.3962	0.688
.*| . |	.*| . |	11	-0.067	-0.143	7.6643	0.743
. |*. |	. | . |	12	0.120	0.051	8.5569	0.740
. | . |	. | . |	13	-0.053	-0.026	8.7388	0.792
**| . |	**| . |	14	-0.271	-0.319	13.656	0.476
. |** |	. |*. |	15	0.198	0.165	16.368	0.358
.*| . |	. | . |	16	-0.125	0.009	17.483	0.355
. | . |	. | . |	17	-0.008	0.046	17.489	0.422
. | . |	. | . |	18	0.030	0.017	17.556	0.485
. |*. |	. |*. |	19	0.095	0.084	18.280	0.504
.*| . |	. | . |	20	-0.114	-0.055	19.379	0.497

Table 4. Augmented Dickey-Fuller Unit Root test on cancer series in Sudan (1967-2010).

ADF Test Statistic	4.43351	1% Critical value*	-3.593
		5% Critical value	-2.932
		10% Critical value	-2.6039

Source: Own construction from study data analysis.

r(k) is the k-th residual autocorrelation and summation is over first s autocorrelations.

(2) Fit versus parsimony: the Schwartz Bayesian Criterion (SBC):

$$SBC = \ln \{RSS/n\} + (p+d+q) \ln (n)/n,$$

where RSS = residual sum of squares, n is sample size, and (p+d+q) the number of parameters.

Having investigated the main feature of cancer data for 1967- 2010 in an attempt to lay the foundation for choice of the appropriate method of fitting a model which best fits the data, and having concluded that the data as in Table 1 is an ARIMA (1, 1) model it is now time for fitting it to the data, that is, its parameters will now be obtained from cancer data.

Model with high adjusted R^2 indicates that the regression line perfectly fits the data, small value of the Akaike information criterion (AIC) is the best model and Durbin–Watson statistic around 2 indicates no autocorrelation in the model, (Table 4).

Cancer is being diagnosed more and more frequently in the Sudan recently because there is lack of practical advice for programme managers and policy-makers on how to advocate, plan and implement effective cancer control programmes.

Incidents of cancer in Sudan for 2011- 2025

Using the developed model, forecasts of incidents expected to occur in future (other things follow similar patterns) are given in Table 5.

Conclusion

Incidents of cancer in Sudan started growing steadily in number since the mid-sixties and are likely to continue growing over the next 15 years (Table 5), if no significant intervention is made by the health authorities. Since the main causes behind such growing numbers remain to be

Table 5. Augmented Dickey-Fuller Unit Root test on first difference cancer series (1967-2010) in Sudan.

ADF Test Statistic	-3.923045	1%	Critical value*	-3.5973
		5%	Critical value	-2.9339
		10%	Critical value	-2.6048

Source: Own construction from study data analysis.

investigated, such interventions are likely to include: increasing the awareness of the importance of early detection and provision of treatment centres etc. However for more concrete solution to the problem, researches on the main cause of cancer in Sudan are inevitable.

REFERENCES

Box GEP, Jenkins GM (1970). Time Series Analysis, Forecasting and Control. Holden Day, San Francisco. pp. 46-87

Box GEP, Jenkins GM, Reinsel GC (1994). Time Series Analysis, Forecasting and Control, 3rd ed., Prentice Hall, Englewood Clifs.

Daoud EH, El Hassan AM, Zak F, Zakova N (1968). Aspects of malignant disease in the Sudan. In Cancer in Africa. P. Clifford et al. Ed. Nairobi, East African Publishing House. pp. 43-50.

El Hassan AM, Lynch JB (1962). Tumors of salivary tissue in the Sudan. Sudan Medical J. 1:3-10.

Federal Ministry of Health (FMOH) (2010). Annual Health Statistics Reports.

Hickey BB (1959). Malignant epithelial tumours in the Sudanese. Annals of the Royal College of Surgeons of England 24:303-322.

Parkin J, Ferlay M, Hamdi C (2003).Cancer in Sudan, in Cancer in Africa Epidemiology and Prevention, International Agency Research on Cancer (IARC) and World Health Organization(WHO) Scientific publication No.153. Lyon, IARC, pp. 40-42.