Research Article

Developments and Applications of Biostatistical Time Series: A Review

Shelton Peiris*[1] and Tim Swartz[2]

[1]*Visiting from School of Mathematics and Statistics, The University of Sydney, Australia*

[2]*Department of Statistics and Actuarial Science, Simon Fraser University, Canada*

*Corresponding author: Shelton Peiri, Visiting from School of Mathematics and Statistics, The University of Sydney, NSW, Australia.

Abstract

Many data sets from medical science are available as time series, especially the records in births, epidemiology, fatal accidents, medical expenses etc, where the information are collected over time. This paper gives a short review of time series methods which have been used in medical research.

Keywords: Time series; Autoregression; Serial correlation; Stationarity; ARMA; Forecasting

Introduction

A time series is a collection of measurements recorded through a suitable time scale. Although this time scale may not be equally spaced, many applications are based on equally spaced time series data. It is known that time series data are generated when a population or an important phenomenon is monitored over time. Many time series analysts use the family of autoregressive integrated moving average (ARIMA) as it enjoys fruitful applications in modelling and forecasting of time series data (ie. serially or autocorrelated data) arise in almost all social sciences. Each member of this family assumes that future values of the series have a clearly defined dynamic parametric relationship which involves both current and past values together with a random noise. A number of extensions to this ARIMA family have been developed to model time series data not following the standard assumptions. A reason for this is that accurate modelling and analysis are useful in practice to estimate potential future observations or forecast values.

Many countries around the world use time series methodology to increase the quality of human life in health, epidemiology, national planning, controlling mortality etc as they a ect their developments. Planning and forecasting of population or migration are also essential for allocation of funds for social services of nations. Therefore, developing good time series models to best suits data sets and use them for accurate forecasting is essential

in applications. This review paper is dedicated and focused on a short review of modeling and forecasting through popular time series models highlighting some potential applications through the R statistical software package.

Popular Time Series Models in Practice

Suppose that we have observed a time series of *n* observations on $\{X_t\}$ following

$$\phi(L)(1-L)^d X_t = \theta(L)e_t, d = 0,1,2....,$$

where L is the lag operator such that $L^j X_t = X_{t-j}(j \ge 0), \{e_t\}$ is a sequence of uncorrelated random variables, $\phi(L) = 1 - \phi L - ... - \phi_p L^p$ is a polynomial of order p and $\theta(L) = 1 + \theta_1 L + ... + \theta_q L^q$ is a polynomial of order q: Since X_t has integrated at order d; (2.1) is known as an autoregressive integrated moving average of order (p,d,q) and is abbreviated by $ARIMA(p,d,q)$: Although it is expressed in a very general form, in many applications, we need integers of (p,d,q) such that $0 \le p,d,q \le 2$, and making ieasy to apply in practice. Stationarity of the series and the orders p; d; q can be identified through the time series plot, autocorrelation (acf) and the partial autocorrelation (pacf) of the data. The time series plot tells whether it needs prior transformations and/or differencing to make it to a stationary series $\{Y_t\}$ Some possible models are:

• If the acf of $\{Y_t\}$ (transformed series) or $\{X_t\}$ (original series) decays very quickly (exponentially), then the time series

is called 'short memory' and use the standard ARMA modelling techniques to and the locallybest possible model.

- If the acf of $\{Y_t\}$ and $\{X_t\}$ decays very slowly (hyperbolically), then the time series is called 'long memory' and use the theory of fractional differencing together ARMA techniques to and the best possible model. The corresponding model is similar to that of (2.1) with a fractional degree of differencing 0 < d < 0:5: This family is known as autoregressive fractionally integrated moving average or ARFIMA(p,d,q).

- Replacing the operator $(1-L)^d X_t, 0<d<0.5$ by $(1-2uL+L^2)^d X_t, |u|<1.$ there is a very general long memory time series known as Gegenbauer ARMA or GARMA(p,d,q; u) with many applications infinance and biostatistics. See for example, Dissanayake et al. (2018) and references therein.

- For certain time series data sets, the acf is not significant for the original data on X_t but the acf of X_t^2 is very significant. Such data can be modelled using the theory on

generalized autoregressive conditional heteroscedastic (GARCH) models. This family of GARCH (p; q) is given by

$$X_t = \sqrt{h_t \varepsilon_t}$$

$$h_t = \omega + \sum_{i=1}^{p} \phi_i h_{t-j}, \omega > 0, \phi_i \geq 0, \theta_j \geq 0,$$

Where $\sum_{i=1}^{\max(p,q)}(\phi_i + \theta_i) < 1, \varepsilon_t$ comprises of a stationary sequence of independent and identically

distributed random variables satisfying zero mean, unit variance and h_t is the conditional variance of X_t given the history; ie. $h_t = Var(X_t | X_{t-1}, X_{t-2}, ...).$

Illustrations

Below we illustrate these three cases in applications. Notice that acf of the last plot shows no serial correlation of data, however, Data2 (or Data2) contains significant acf values (Figure 1).

Now we look at an application of ARIMA modelling using a biostatistical time series data set.
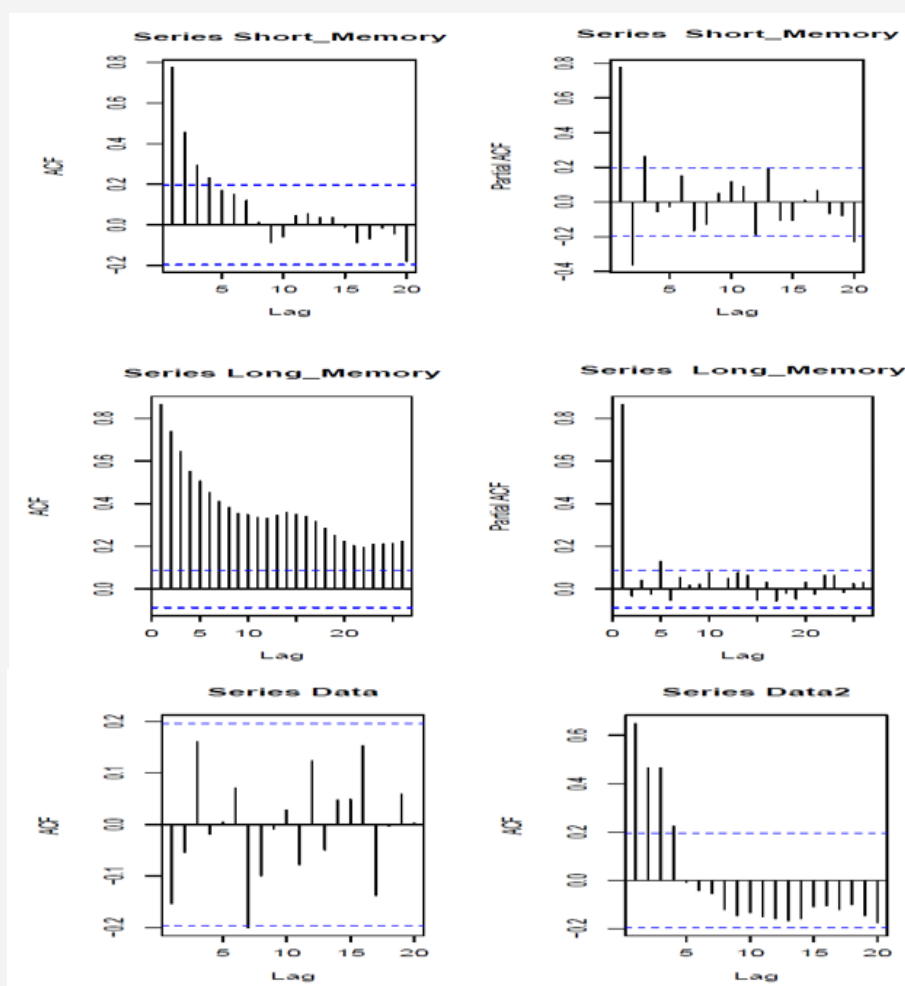

Figure 1

ARIMA Modelling in Biostatistics

This section considers an application from annual incidences of stomach cancer in Australia from 1982 to 2015 (available in

Cancer data in Australia https://www.aihw.gov.au/) to illustrate the procedure. The time series plot of data shows a slight upward trend. We conduct the analysis based on both the original and lag

1 differenced data and select the best possible model from a pool of all potential models. Time series plots of cancer cases (original data) and Differenced data (lag one differenced data) together with corresponding acf and partial pacf are given below: (Figure 2).
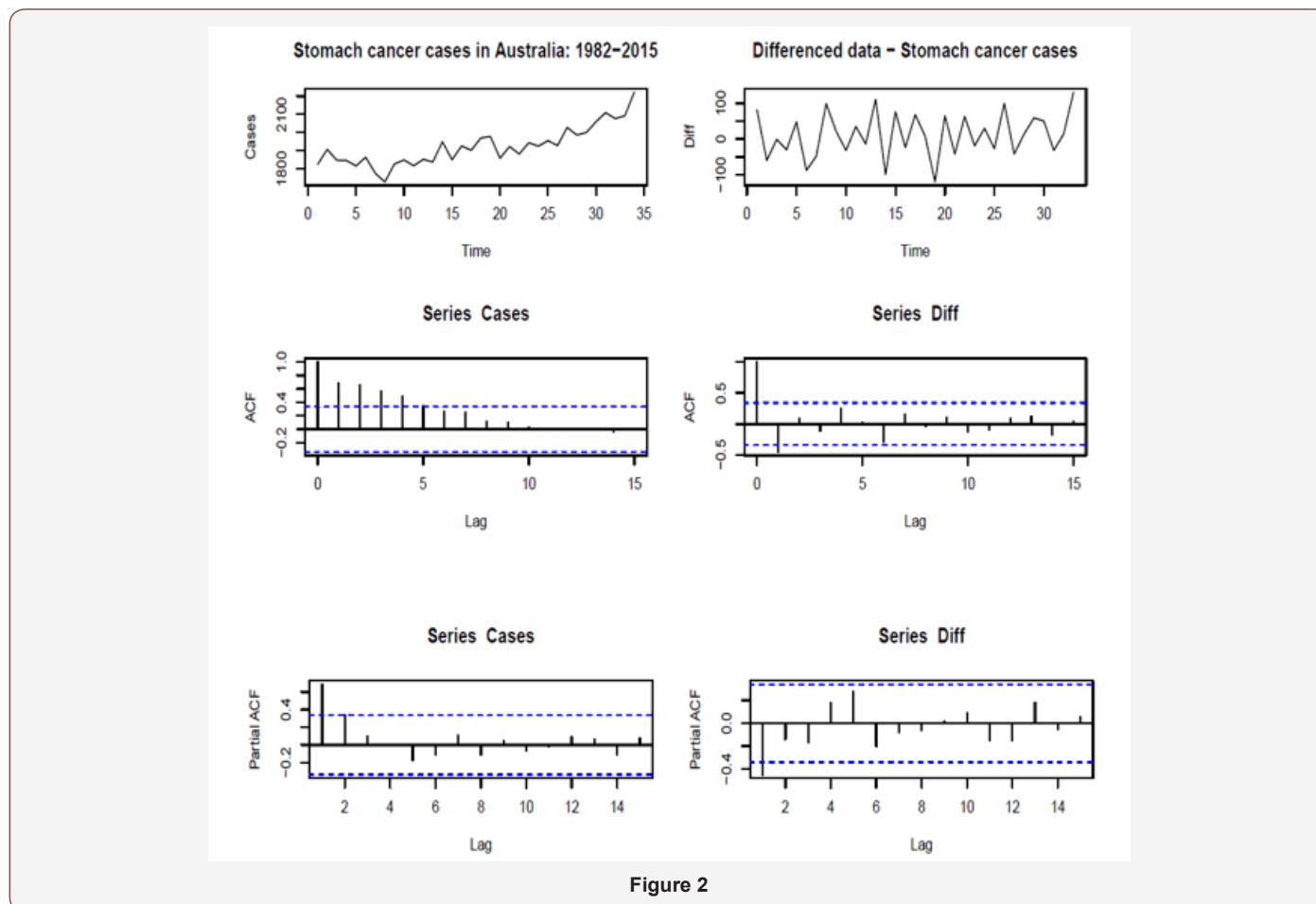

Figure 2

Based on the above plots, suggest the following ARMA models: Table 1

Table 1:

Data: Cancer cases	Di erenced Data
AR(1)	AR(1)
AR(2)	MA(1)
	ARMA(1,1)

Estimating each model using R give the following summary (values in parentheses are the corresponding estimated standard errors) Table 2:

Table 2:

Data	Model	Estimates	AIC
Cases	AR(1)	ar1=0.8817 (0.0960)	384.81
		intercept=1953.1076 (77.7157)	
	AR(2)	No parameter estimates	-
Di er-enced	AR(1)	ar1=-0.5138 (0.1607)	363.48
		intercept=10.0461 (6.3119)	
	MA(1)	ma1= -0.5330 (0.1407)	362.87
		intercept=9.6618 (4.5416)	
	ARMA(1,1)	ar1= -0.2691 (0.2782)	364.07
		ma1=-0.3583 (0.2475)	
		intercept=9.5623 (4.7969)	

This tells us that the differenced MA(1) model for the differenced data is the best among those have been suggested since it has the smallest aic value. Let $\{X_t\}$ be the original series and $Y_t = X_t - X_{t-1}$ represents the differenced series. Therefore, the best fitted model for Y_t is

$$Y_t = 9.6618 + e_t - 0.5330 e_{t-1}$$

where $\{e_t\}$ is the associated noise.

Further applications of ARIMA modelling in biostatistical time series can be found in Cox and Solomon (1988), Zeger et al. (2004), Zeger et al. (2004), Perera et al. (2008), Zollar et al. (2016), Siskina and Siaulys (2016), Yousefzadeh-Chabok et al. (2016), Ferenci (2017), Fried et al. (2017), Xie (2017), Pavareh et al. (2018),Dissanayake et al. (2018), Earnest (2019) and references theirin.

Acknowledgement

None.

Conflict of Interest

No conflict of interest.

References

1. Cox DR, Solomon PJ (1988) On testing for serial correlation in large numbers of small samples. Biometrika 75: 145-148.

2. Dissanayake GS, Peiris MS, Proietti T (2018) Fractionally dierenced Gegenbauer processes with long memory: a review. Statistical Science 3: 413-426.

3. Earnest A, Evans S, Sampurno F, Millar J (2019) Forecasting annual incidence and mortality rate for prostate cancer in Australia until 2022 using autoregressive integrated moving average (ARIMA) models, BMJ Open.

4. Ferenci T (2017) Biomedical applications of time series analysis, IEEE 30th Jubilee Neumann Colloquium: 83-84.

5. Mishra AK, Sahanaa C, Manikandan M (2019) Forecasting Indian infant mortality rate: An application of autoregressive integrated moving average model. J Family Community Med 26(2): 123-126.

6. Fried R, Abbas S, Borowski M, Imho M (2017) Online Analysis of Medical Time Series. Annu Rev Stat Appl 4: 169-188.

7. Pavareh M, Karimi A, Rezaei S, Woldemichael A, Nili S, et al. (2018) Assessment and prediction of road accident injuries trend using time-series models in Kurdistan, Burns & Trauma: 6-9.

8. Perera DI, Peiris S, Robinson J, Weber NC (2008) The empirical saddlepoint method applied to testing for serial correlation in panel time series data. Statistics and Probability Letters 78: 2876-2882.

9. Siskina N, Siaulys J (2016) ARMA models for mortality forecast. Lithuanian Journal of Statistics 52(1): 31-44.

10. Xie L (2017) Time Series Analysis and Prediction on Cance rIncidence Rates. Journal of Medical Discovery 2(3).

11. Yousefzadeh-Chabok S, Ranjbar Taklimie F, Malekpouri R, Razzaghi A (2016) A Time Series Model for Assessing the Trend and Forecasting the Road Tra c Accident Mortality. Arch Trauma Res 5(3): e36570.

12. Zeger Scott L, Irizarry, Rafael A, Peng, Roger D (2004) On Time Series Analysis of Public Health and Biomedical Data (September 2004). Johns Hopkins University, Dept of Biostatistics Working Papers. Working Paper 54.

13. Zeger Scott L, Irizarry Rafael A, Peng Roger D (2006) On Time Series Analysis of Public Health and Biomedical Data. Annu Rev Public Health 79: 27-57.

14. Zolala F, Haghdoost A, Ahmadijouybari T, Salari A, Bahrampous A, et al. (2016) Forecasting the Trend of Trac Accident Mortality in West Iran, Health Scope 5(3).