Research Article

Time Series Analysis and Prediction on Cancer Incidence Rates

Liming Xie[1,*]

[1]Department of Statistics, North Dakota State University, 1230 Albrecht Blvd, Fargo, ND 58108

Abstract

Background Time series is a useful analysis method for any data. Many researchers used it to analyze their observations to obtain the results they want. The aim of this article is to apply some time series techniques such as autoregressive integrated moving average (ARIMA) model to analyze data of cancer incidence rates (CIR) and its residuals in the United States for 1975-2014.

Objective The aim of this article is to apply some time series techniques to analyze data of CIR and its residuals in the United States to find fit model and make reasonably forecast the trends for 2015 to 2020.

Methods Author Applies Autoregressive integrated moving average (ARIMA) model and SARIMA model (Seasonal ARIMA models) to diagnose CIR data. Using R version 3.4.0 (2017-04-21). The total observations are160 groups with 40 years and frequency=4) that include White males and females, Black males and females in the United States for 1975-2014.

Results Although there are not big differences in 160 observations between AIC and BIG, the best model is still confirmed by ARIMA (1,1,2) (1,1,1) [4]. The model function is the following: (1+0.9751B) (1-0.0269B^2) (1+0.6751B^{12}) X_t (1-0.3360B^{12}) (1+0.8033B^{24}) $W_t.W_t \overset{iid}{\sim}$N (0, 0.0002226). SARIMA models are (1,1,2) and model (1,1,1). Both are fit for normality. ACFs of residuals are not deviated from the model assumptions. Q-statistic p-value is fit for ARIMA (1,1,2) and ARIMA (1,1,1) models.

Conclusion When using ARIMA model and SARIMA scripts for 160 observations of data, author finds some significant points in time series. The results indicate that the trends of CIR for 2015-2020 is downward by 300-550 per 100,000 population in the United States each year.

Keywords: *CIR, Time Series, ARIMA, SARIMA, Forecasting, Trends*

How to cite: L Xie, Time Series Analysis and Prediction on Cancer Incidence Rates. J Med Discov (2017); 2(3): jmd17030; DOI:10.24262/jmd.2.3.17030; Received July 4th,2017, Revised August 20th,2017, Accepted August 29th,2017, Published September 5th,2017.

Introduction

Cancer is one of the most cause of death in the world. It is more and more menace to human being. For many years

* Correspondence: Liming XieDepartment of Statistics, North Dakota State University, 1230 Albrecht Blvd, Fargo, ND 58108

a lot of physicians and researchers have been doing unremitting efforts and working. Many institutions take a lot of statistical surveys and research. National Cancer Institution (NCI) at American National Institution of Health(NIH) usually reports and provides annual data and statistical reviews. SEER, Surveillance, Epidemiology and End Results Program from NIH offers some the most recent incidence, mortality, prevalence, survivals statistical data. "SEER Cancer Statistics Reviews in 1975-2014" in the United States includes some brows the table and figures that indicate Delay Adjusted Incidence, incidence by race/ethnicity, rates for sex, age, and year of diagnosis of major cancers. Times series is a series of data by graphs, list or other forms. It can predict ahead data trend of the series. The author tries to use this review to statistically analyze CIR in the United States from 1975 through 2014, and then predict it for 2015-2020. The author expects to find its epidemiolocal characteristics and make the guide to prevent or avoid some cancer occurrence.

CIR is to calculate the numbers of getting a or more particular type of cancer cases per 100,000 people from a location yearly. Its formula is "Incidence rate = (New cancers / Population) × 100,000**"** [1]**.** In this formula, "New cancers" denotes the number of catching new cancer. "Population" denotes the number of cancers per 100,000 people at risk. Many researchers found some characteristics, such as, in Asia and Africa both breast cancer have been a faster trend to rise for incidence rates than other areas [2]. However, some researchers found that the situation was depended on aged groups of breast cancer: trends for women aged 45 to 54 years in the United States have continued to be dropped since 1980, aged 45-74 years a little increased trend in the1980s but it declined to a lower level in the 1990s; the trend of getting breast cancer for aged 45 to 74 years in some European countries slightly increased before 1980 but it decreased later [3]. However, in some developmental countries some cancer incidence rates still increased, such as breast cancer was "2.62% per year", liver cancer was "4.43 % per year" [4]. In the United States, some researchers found that the incidence rate of sex difference for some cancers was significant, for example, Esophagus cancer was "-27.9%" for American males, and "-17.6%" for American females, but both trends were increased: lung cancer, "133.4%" for American males, "107.7%"; in bladder cancer both were significant, "20.9%" for American males, "-25.7%" [5]. Some male cancer, such as prostate cancer, also increased "22.8%" [5]. In Asia, some researchers surveyed that "Korean men have a lifetime cumulative incidence rate of less than 10% for five specific cancers (stomach, lung, liver, colon, and prostate)." [6]. CIRs of some cancers in China were high: for examples, stomach, esophageal and colorectal cancers from 2003 to 2012 were "95.56 %, 45.60%, 14.62%", respectively [7]. In CIR of gynecological tumors, cervical cancer in China was lower than other countries and areas [8]. One in Japan and Korea were "12%, 21.1%, 7.3 % in China". In three countries, China, Japan and Korea, for corpus cancer, CIRs were "2.2%, 9.12%", respectively; for ovary cancer, there were "3.4%, 10.1 %, 6.6%", respectively; "20.1%, 49.6%, 23.5%" for breast cancer [9]. Therefore, CIR is very important to study the epidemiology of cancers and prevent them in any areas. It has even key role on guideline to treat a various of cancers.

Materials and Methods

Data Source

Data comes from National Cancer Institute(NIH). It consists of male and female groups for the White and the Black in the United States who is 160 groups among all cancers, each year includes 4 groups and the total of 40 years during 1975-2014. In data type, it is used "SEER incidence (Surveillance, Epidemiology, and End Results

Program from NIH)"; in statistical type, "Age-Adjusted Rates" is used, in cancer site, it is by "All Cancer Sites Combined". The raw data is as follows:

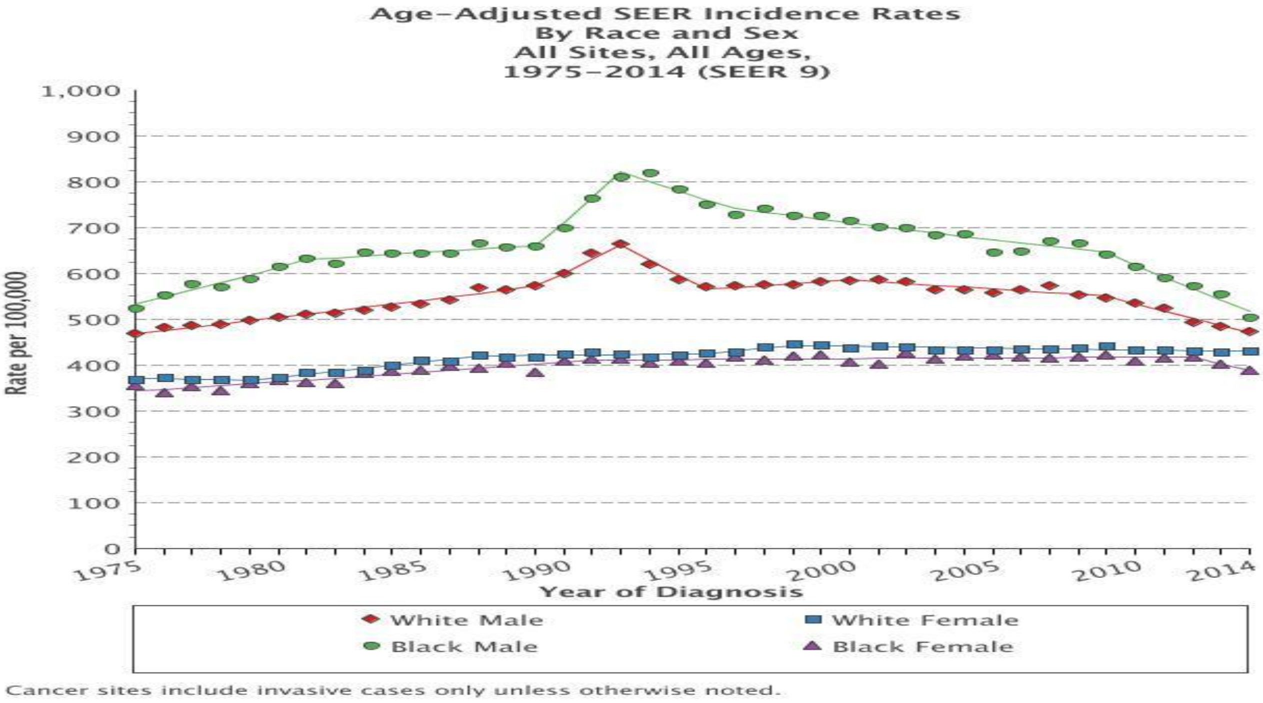


Figure 1. Table of Cancer Incidence Rates in the United States in 1975-2014.

For Age-Adjusted Rates, it includes each of the 5 age groups for the same geographic, the white and the black in the United States. The data provide information from the aged groups: ages <20, ages 20-49, ages 50-64, ages 65-74, and 75+. A crude incidence rate is the number of new cancers a specific site/type occurring in a specific population during a year. Usually NIH uses the following expression:

$$\text{Crude rate} = \frac{count}{population} \times 100{,}000^{[11]}.$$

Statistical Analysis

In this paper the author utilizes R version 3.4.0 (2017-04-21) to analyze CIR data and tried to use ARIMA (the Autoregressive Integrated Moving Average) Model. Today, ARIMA is often used to analyze and explain most of data, and forecast the future values of the standard errors. The model is as follows:

$$X_t = \theta_1 x_{t-1} + \theta_2 x_{t-2} + \cdots + \theta_p x_{t-p} + w_t$$

Here,X_t is stationary, θ_1, θ_2, …, θ_p are constants, w_t is Gaussian white noise series with mean zero and variance σ_w^2[12].

Firstly, the author uses STL (Seasonal and Trend decomposition using Loess) function to see if the time series of CIR data exist in seasonality (Figure 2). In differencing, since original data is not stationary, the author applied the differencing technique for onetime. The formula is:

$$Y_t' = y_t - y_{t-1} \ \big(\text{1stDifferencing}(\text{d} = 1)\big).$$

Here, the difference data between consecutive observations is calculated. To make a series stationary on variance, the author tries to apply transforming the original series by log transform, after first time differencing data (Figure 3).

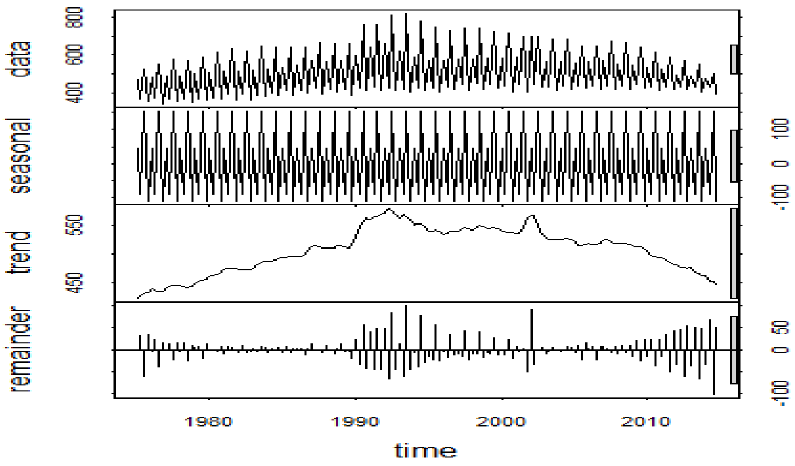


Figure 2. STL plot. Trend showed downward and severe seasonality. Time series is not stationary on variance.

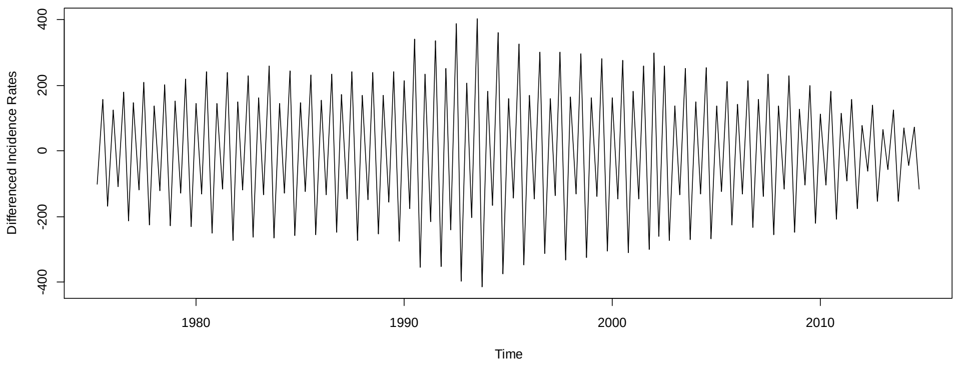


Figure 3. First difference of the time series of the CIR data. It still showed trend but a little better than downward original time series.

In Autocorrelation plot it is used by ACF (The autocorrelation function that tests the linear predict ability of the series) and PACF (Partial autocorrelation function that may have partial correlative, lagged and significant values to a time series) to determine whether a series is stationary, and these plots are helpful to pick up the order parameters for ARIMA model (Figure 4). ACF plots display correlation between a series and its lags; PACF plots display between a variable and its lags that is not explained by previous lag. In choosing the model order, the author used a criterion to determine the data order of no seasonal ARIMA model is the Akaike Information Criterion (AIC), that is: AIC=−2 log(L)+2(p+q+k+1), here, L is the likelihood of the data, p is the order of the autoregressive part and q is the order of the moving average part. The parameter k in this criterion is defined as the number of parameters in the model being fitted to the data. Also, the Bayesian Information Criterion (BIC) is used: BIC= AIC +(log(T)-2) (p+q+k+1). This is to minimize the AIC, AICc or BIC values for a good model. Because a lower value of one of these criterions for a range of models is the best suit for the data (See the following output from R).

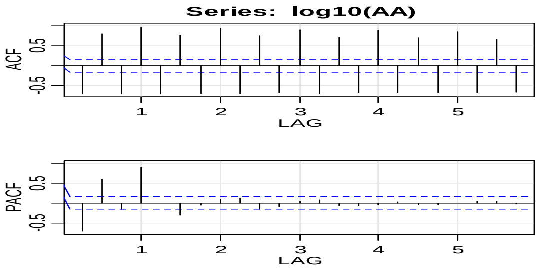


Figure 4. After log of the data, the ACF seems seasonality, the series seems to be stationary on variance.

To forecast the values of data in the future years by using ARIMA, the author used 95% prediction interval for x_{n+m}^{n}, that is, for the assumption of normally distributed errors, a 95% prediction interval for x_{n+m}^{n}, the future value of the series at time n+m,

$$x_{n+m}^{n} \pm 1.96\sqrt{\hat{\sigma}_w^2 \sum_{j=0}^{m-1} \Psi_j^2}.$$

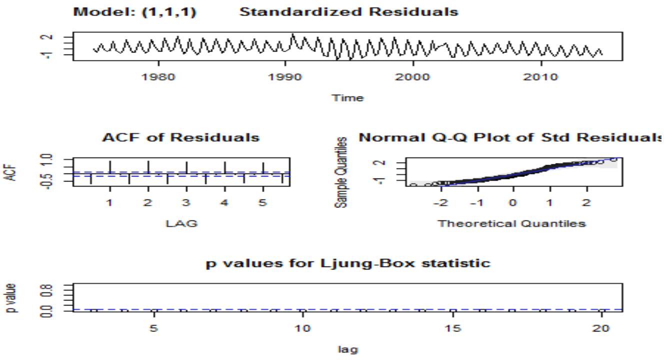


Figure 5. SARIMA output of Model (1,1,2). The plot of Standardized Residuals displays no significant pattern.

In this article next 6 years is predicted after year 2014. To diagnose series of the cancer incidence rate, the author applied **sarima** scripts **(**Seasonal ARIMA, a regular pattern with changes that usually appeared reduplicatively in S time periods**)**. For example, **sarima** (log(data)), 0,1,1, no constant =TRUE). The main models were an AR (1) and an MA (1), such as ARIMA (0,1,1) and ARIMA (1,1,1) (See Figure 5 and Figure 6). AIC prefers the MA (2) and BIC prefers the AR (1) model. It is often the case that the BIC will select a model of smaller order than the AIC or AICc. It would not be unreasonable in this case to retain the AR

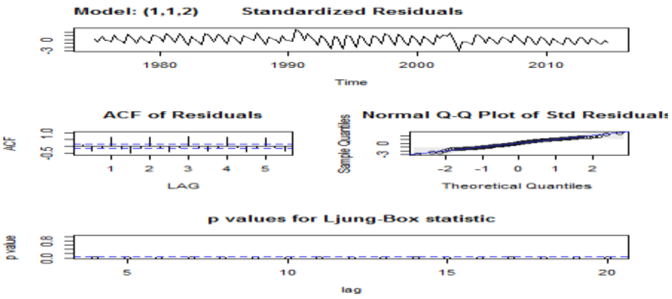


Figure 6. SARIMA of model (1,1,1). Standardized Residuals displays same as the model (1,1,2).

(1) because pure autoregressive models are easier to work with.

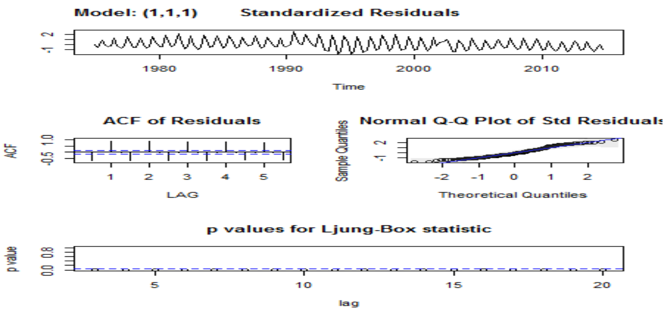


Figure 7. SARIMA of model (1,1,1).,the plot of Standardized Residuals

Displays same as the model (1,1,2).

Results

For model identification, ARIMA model is suitable to data of CIR. the author used method of STL scripts to overview original data and see if it exists in stationary. Otherwise, make some transform the variables or differencing the series, etc. Since the stationary series is the one whose values change over time only under a constant mean and variance. So, we check it by some plots or graphs from the data. When using STL function, the author found that there is downward trend and severe seasonality for the original data. Therefore, in general, first difference to the original data is applied to have data be reliable prediction. Differencing is an excellent way to transform a non-stationary series to a stationary one. After first difference to time series of CIR data, the process still did not remove trend but a little better than previous one. So, it is necessary to transform data by logarithm to make a series stationary on variance, that is use the following mathematical formula: $X_t^{new} = \log_{10} X_t$, where X_t is time series of CIR data. Since using the original CIR data without differencing, the series looks like non-stationary. This requires to differencing log10 of original data. Thus, the series has stationary on both mean and variance. However, plots of ACF and PACF showed a lot of spikes at dotted horizonal lines. So, no random at the residuals. Therefore, the estimation of parameter of the model and diagnostics on it are the most important things.

Both AIC and BIC are used to whether the models are fit criterion. AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data. BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup. A lower AIC value means that a model usually is probably be closer to truth. A lower BIC value means that a model might to be the true model. There is a little difference in both application: It is usually pick up AIC if having a big model; BIC is used to smaller model. Since observations for this article n=160 is not big, both are not significant difference. The author confirmed that the best model is ARIMA (1,1,2) (1,1,1) [4]. The model function is the following:

$$(1+0.9751\mathrm{B})\ (1\text{-}0.0269B^{2})(1+0.6751B^{12})X_t\ \ (1\text{-}0.3360B^{12})\ (1+0.8033B^{24})\ \ W_t.W_t \overset{iid}{\sim} \mathrm{N}\ (0,\ 0.0002226).$$

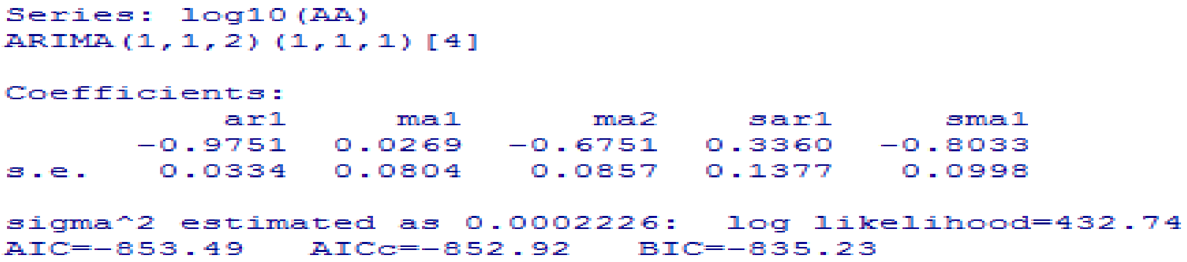
On the other hand, SARIMA model is used to diagnose the models of the CIR data. Plots display that both model (1,1,2) and model (1,1,1) are fit for normality. ACFs of residuals are not deviated from the model assumptions. Q-statistic p-value is fit for ARIMA (1,1,2) and ARIMA (1,1,1) models.

Regarding forecasting CIR data and analysis, the author used ARIMA (1,1,2) (1,1,1) [4] to predict the values ahead 6 years after 2014, that is, from 2015 to 2020. The output showed that the CIR values for these years are downward trends and should be the range of 300-550 per 100,000 population in the United States.

Discussion

ARIMA model is one of most popular models for time series. For many years, people apply it to analyze a various of data and forecast the values of residuals. In fact, they obtained the cohort effect and achieved the goals they expect. Some researchers think that the ARIMA model has more flexible modeling for trends and some parameters [20]. Some Chinese scholars used ARIMA model to analyze the incidence of hemorrhagic fever with Renal Syndrome, and then made short-term forecasts [21]. One researcher used it to identify some mortality rates [22]. Rather, some researchers applied this model to link with general regression neural network model and made both to analyze and forecast hepatitis incidence in some countries [23]. Other researchers even think that it is a good method that does not require to select a priori the value of any parameter for any time series [24]. These cases and applications displayed that ARIMA model was a useful technique and very strength to analyze and forecast the values of standard errors.

ARIMA (p, d, q)× (P, D, Q) S, where p is non-seasonal AR order, d is non-seasonal differencing, q is non-seasonal MA order, P is seasonal AR order, D is seasonal differencing, Q is seasonal MA order, and S is timer span of repeating seasonal pattern. The output of Ris the following:



```
Series: log10(AA)
ARIMA(1,1,2)(1,1,1)[4]

Coefficients:
          ar1     ma1      ma2    sar1     sma1
      -0.9751  0.0269  -0.6751  0.3360  -0.8033
s.e.   0.0334  0.0804   0.0857  0.1377   0.0998

sigma^2 estimated as 0.0002226:  log likelihood=432.74
AIC=-853.49   AICc=-852.92   BIC=-835.23
```

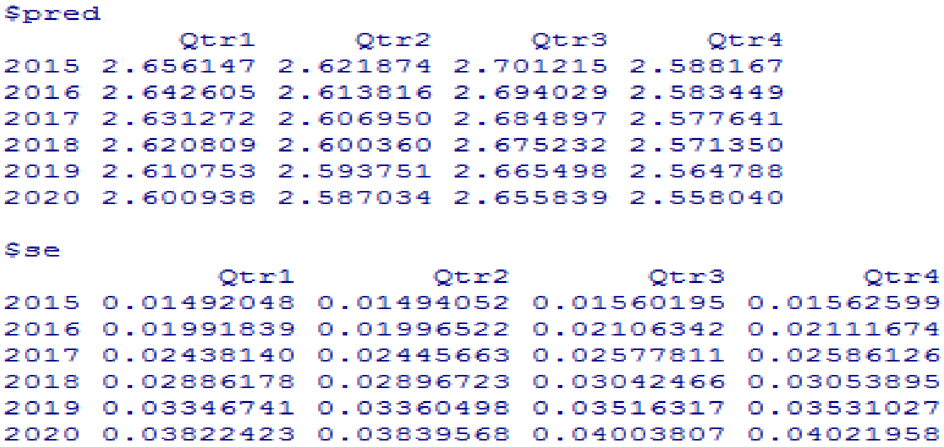
The above result shows ARIMA (1,1,2) (1,1,1) [4].

Meanwhile, the Model function is written as:

$$(1+0.9751B) \quad (1-0.0269 \ B^2 \ ) \quad (1+0.6751 \ B^{12} \ ) \quad X_t$$

$$(1-0.3360B^{12}) \ (1+0.8033B^{24}) \ W_t.W_t \overset{iid}{\sim} \mathrm{N} \ (0, 0.0002226).$$

On the other hand, the output of the series showed the standard errors for ahead 6 years.



```
$pred
         Qtr1     Qtr2     Qtr3     Qtr4
2015 2.656147 2.621874 2.701215 2.588167
2016 2.642605 2.613816 2.694029 2.583449
2017 2.631272 2.606950 2.684897 2.577641
2018 2.620809 2.600360 2.675232 2.571350
2019 2.610753 2.593751 2.665498 2.564788
2020 2.600938 2.587034 2.655839 2.558040

$se
           Qtr1       Qtr2       Qtr3       Qtr4
2015 0.01492048 0.01494052 0.01560195 0.01562599
2016 0.01991839 0.01996522 0.02106342 0.02111674
2017 0.02438140 0.02445663 0.02577811 0.02586126
2018 0.02886178 0.02896723 0.03042466 0.03053895
2019 0.03346741 0.03360498 0.03516317 0.03531027
2020 0.03822423 0.03839568 0.04003807 0.04021958
```

se= Standard Errors; Qtr1=White males, Qtr2=White females, Qtr3=Black males, Qtr4=Black females.

To confirm the mode fitting, the author applied **sarima** technique. MA (2) and AR (1). It includes the formula: “ $e_t = \frac{x_t - \hat{x}_t^{t-1}}{\hat{P}_t^{t-1}}$, where $\hat{x}_t^{t-1}$ is the one-step-ahead prediction of $x_t$based on the fitted model and $\hat{P}_t^{t-1}$is the estimated one-step-ahead error variance [12] [13].

All values are within 3 standard deviations in magnitude, although there are some outliers. The ACF of residuals

seems to be deviated from the model assumptions. The normal Q-Q plots of the residuals do not show deviated from normality at the tails.

Finally, the plot of forecasting ahead 6 years for 2015-2020 showed that the trend of these years might to be downward and the range of them is below between 300 and 550 per 100,000 population in the United States every year (Figure 6).

We plot ACF and PACF of the residuals of the best fit ARIMA model. The model of the data is ARIMA (1,1,2)(1,1,1) [4].

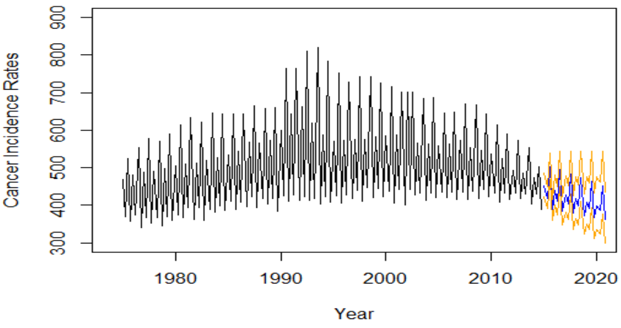


Figure 8. Forecasting ahead 6 years after 2014. The plot showed that, during 2015-2020, the CIR should be generally decreasing trend and between 300-550 per 100,000 people each year.

CIR is an important index of cancer that records new cancers of some region or type incidence in a specific population at risk during specific time. If we understand some probability and statistical indicators that estimate and predict on CIR. These will have far-reaching implications for medicine, epidemiology and treatment of cancers.

Competing Interests

None

Acknowledgements

The authors are pleased to acknowledge the assistance from Shanshan Hu offered in the analysis.

References

1. NIH: Surveillance, Epidemiology, and End Results Program; Defining cancer statistics; Cancer Incidence Rates. 2017.
2. A. Sasco. Epidemiology of breast cancer: An environmental disease, Apmis.2001;109 (5) 321-32.
3. Bircan E, Muhammed A, Dorota M G, et al. Using Functional Data Analysis Models to Estimate Future Time Trends in Age-Specific Breast Cancer Mortality for the United States and England-Wales. J Epidemiology. 2010;20(2): 159-65.
4. Michael P, Gladys H, Lissette R, et al. Trend Analysis of Cancer Mortality and Incidence in Panama, Using Joinpoint Regression Analysis. Medicine. June 2015; 94(24).
5. Edwin S, and Arthur I H. Major Trends in Cancer:25 Year Survey. CA: A Cancer J for Clinicians. Dec 2008; 25(1).
6. Soyeun K, Dong W S, Hyumg K Y, et al. Public Perceptions on Cancer Incidence and Survival: A Nation-wide Survey in Korea. Cancer Res Treat. Apr 2016; 48(2): 775-88.
7. Li C Y, Ye Y C, Liang G Y, et al. Cancer Incidence and Mortality Survey in Wuwei, Gansu Province, Northwestern China from 2003 to 2012: A Retrospective Population-based Study. Chin Med J(Engl). May 2016;129(6): 636-44.
8. Shi J F, Canfell K, Lew J B, et al. The burden of cervical cancer in China: synthesis of the evidence. Int. J Cancer. 2012; 130: 641-52.
9. Kim K, Zang R G, Choi S C, et al. Current Status of Gynecological Cancer in China. J GynecolOncol. June 2009;20(2): 72-76.
10. NIH: Surveillance, Epidemiology, and End Results Program: Interactive Tools; Fast Stats, Race and Sex.2017.
11. NIH: Surveillance, Epidemiology, and End Results Program: For Researchers; Fast Stats, SEER*Stat Tutorials. 2017.
12. Shumway R H, and Stoffer D S. Time Series Analysis and Its Applications with R examples (3rd edition). Springer. 2011;84-85.
13. Shumway R H, and Stoffer D S. Time Series Analysis and Its Applications with R examples (3rd edition, Blue Printing). Springer. 2015.
14. Farah Y, and Sidra Z. Functional Times Series Models to Estimate Future Age-Specific Breast Cancer Incidence Rates for Women in Karachi, Pakistan. J of Health Science 2. 2014; 213-21.
15. Kehav P. Pokhrel and Chris P. T. Forecasting Age-Specific Brain Cancer Mortality Rates Using Functional Data Analysis Models. Adv in Epidemiology. 2015.
16. Wah W, Das S, and Earnest A, et al. Time Series Analysis of Demographic and Temporal trends of Tuberculosis in Singapore. BMC Public Health. 2014; 14:1121.
17. Kitagawa G. Introduction to Time Series Modeling. Chapman&Hall, Boca Raton, FL. 2010.
18. Lutkepohl H. New Introduction to Multiple Times Series Analysis. Springer. 2005.
19. Montgomery D C, Jennings C L, and Kulahci M. Introduction to Times Analysis and Forecasting. John Wiley & Sons. 2008.
20. Shibuya K, Inoue M, and Lopez A D. Statistical Modeling and Projections of Lung Cancer Mortality in 4 Industrialized Countries. Int. J Cancer. 2005;117: 476-85.
21. Li Q, Guo NN, Han ZY et al. Application of an Autoregressive Integrated Moving Average Model for Predicting the Incidence of Hemorrhagic Fever with Renal Syndrome. Am J Trop Med Hyg. Aug 2012; 87(2): 364-70.
22. Kis M. Time Series on Analysing Mortality Rates and Acute Childhood Lymphoid Leukaemia. Connecting Medical Informatics and Bio-Informatics, IOS Press. 2005.
23. Wei W, Jiang JJ, Liang H, et al. Application of a Combine Model with Autoregressive Integrate Moving Average(ARIMA) and Generalized Regressive Neural Network (GRNN) in Forecasting Hepatitis Incidence in Heng County, China. Plos One. 2016; 11(6): e0156768.
24. Langat A, Orwa G, Koima J. Cancer Cases in Kenya; Forecasting Incidents Using Box & Jenkins Arima Model. Bio Statistical Informatics. April 2017; 2 (2): 37-48.