Analyzing and Forecasting HIV DataUsing HybridTime Series Models

Liming Xie

Department of Statistics, North Dakota State University

Abstract

In real work, we often confront complete linear and nonlinear time series data. But some time series are not pure linear and nonlinear, or complicatedone, we need apply two or more models to analyze and predict them. It is necessary to explore and find some novel time series hybrid methods to solve it. Human Immunodeficiency and Virus (HIV) is one of intractable and trouble diseases in the world. Thus, the author of this article wants to analyze and probe into some novel time series methods to get breaking breach in the epidemiology that find some rules in the incidence, distribution, pathogen, and control of HIV in a population. In this article, to find the best model, auto.arima function is applied to the original time series data to determine autoregressive integrated moving average, ARIMA(0,0,0); ARIMA and generalized autoregressive conditional heteroskedasticity (GARCH), that is, ARIMA-GARCH (1,1) model is used to analyzenumbers of people living with HIV for the data of HIVin the world such some important parameters as mu, ar1, ar2, omega, alpha 1, or beta 1 and some specific tests, for example, Jarque-Bera Test, Shapiro-Wilk Test, Ljung-Box Test, etc.Using ARIMA (0, 0, 0) and SARIMA (0,2),seasonal ARIMA, to predict the future values and trends after 2015. Both suggest identical results.

Keywords HIV, hybrid, ARIMA, GARCH, SARIMA, Forecast

1. Introduction

Background

HIV is an intractable disease that menaced human health. Its incidence, distribution and how to control are difficult problems until today. Since there are many different statistical reports or results from different institutions or organizations regarding the numbers of people living with HIV, the author of this article uses the living people with HIV 2000-2015 fromauthoritative organization, world health organization (WHO). This article is to seek a breakthrough in HIV survival. In particular, studyingliving population with HIV in the world is crucial to its prevention, and treatment.ARIMA is a well-known time series model to analyze the data and predict the future values and trends. However, some time series data are hard to be explain by ARIMA. It is needed to hybrid some advanced models, such as ARIMA-GARCH, to assess the standard errors, residual tests, and information criterion statistics. It emphasizes to obtain important parameters such as mu, ar1, ar2, omega, alpha 1, or beta 1, etc. Also, some tests such as Jarque-Bera Test, Shapiro-Wilk Test, Ljung-Box Test. These results will determine the model of HIV data. Those parameters help to build ARIMA (2, 0) -GARCH (1, 0). After that, comparing ARIMA (0, 0, 0) (0, 1, 1) with SARIMA.for function to forecast the future values and trends of the data so as to predict accurately.

The author email address:limingxie2018@gmail.com.

Objective

a. Studying deeply the properties of various of time series models and analyzing some similarities and different points among different models to apply them the complicated data so as tosolve the complicated data.
b. Utilizing statistical techniques and methods, such as time series models to find and explore some approach dealing with HIV or prospective thoughts to analysis on the epidemiology of HIV.

Dataset

The data including all ages from WHO is collected by national health departments each country and some government authoritative organizations each five year from 2000 through 2015, for example, "Estimated number of people living with HIV 2015", number of 2010, 2005, and 2000, etc.Estimated numbers of people living with HIV defines that a country is the average mean from the minimum through the maximum, or approximate value. For example, in 2000, the average value was 2400 range from 2200 to 2600 in Algeria. To facilitate the statistics and calculation, the author shrinks 10,000 timeseach data. For example, in 2000, the average value was 2,000,000 in India. It became 200 units when data was analyzed. Some countries cannot report their data such as Albania, Antigua and Barbuda, Bhutan, Canada, etc. It is possible because they have either poor statistical techniques and not report to WHO, or not finding HIV populations or less people, etc. The author of this article extracts countries where have specific statistical numbers and excluding those without data no matter what are reasons. The total countries used in data analysis is 105 which are 420 observations during 2000, 2005, 2010, and 2015.

2. Proposed Methodology

2.1 GARCH Model

A. Definition

***GARCH*(Generalized ARCH models)**

The GARCH (p, q) process is a generalized process of ARCH (p). its formula is as follows:

$h_t = \theta_0 + \sum_{i=1}^{s} \theta_i \varepsilon_{k-i}^2 + \sum_{j=1}^{t} \gamma_j \varepsilon_{k-j}^2$,
where $\theta_0 > 0$ and $\theta_i, \gamma_j \geq 0$, j=1,2, …
The conditional variance may be

$$E_{k-1}\varepsilon_k^2 = \theta_0 + \sum_{i=1}^{s} \theta_i \varepsilon_{k-i}^2 + \sum_{j=1}^{t} \gamma_j h_{t-j} \quad (1)$$

B. Models

Let$y_t$be the value of an object at a time t, then we get the return a series, $x_t$which is the object at time t:

$$x_t = \frac{y_t - y_{t-1}}{y_{t-1}}.$$

where y_t implies (1+x_t) y_{t-1}.
Thus, when the return denotes a small percentage change, we get the following function:

$\square[\log y_t] = x_t$,
where ∇ [log y_t] expresses the return.
In general, the return is not a constant conditional variance. There is dependent of sudden bursts of variability in a return on the series own past.
For GARCH there are two ARMA models below:

ARMA (p, q): $y_t = \theta_0 + \sum_{i=1}^{s} \theta_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{t} \gamma_j \varepsilon_{t-j}$
ARMA (r_p, r_q): $\varepsilon_k = W_q \sqrt{h_t}$, $h_t = \theta_0 + \sum_{j=1}^{r_p} \gamma_j h_{k-j} + \sum_{i=1}^{r_q} \theta_i \varepsilon_{k-i}^2$

For conditional mean is search over alternative ARIMA (p, q) models by varying p and q parameters, and identifying the optimum model using Schawart Information Criterion. The results may model as a MA (1) process. that is,

$T_1 = (1 + \tau L)\, \varepsilon_t$, (2) where τ is a parameter of MA (1), L is the lag operator, and are innovations with zero mean but potential subject to conditional heteroscedasticity. Hence, given the short memory of the MA (1,1) process, the effect of random shocks dissipates after one period. Shumway and Stoffer [2] provided the simplest ARCH model, the ARCH (1), models the return as
$x_t = \theta_t \varepsilon_t$ (3)

$\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2$, (4)where x_t is the return of stable, andε_t is standard Gaussian white noise; that is, ε_t ~iid N (0,1). Also, imposing some constrains on the model parameters so as to get desirable properties.

For ARCH (1) models return with non-constant conditional variance, its conditional variance is relayed on the previous return. If the conditional distribution of $d_t$given d_{t-1}, Gaussian is as follow as:

$x_t \mid x_{t-1} \sim$ N (0, $\alpha_0 + \alpha_1 x_t^2$) (5)

If equation (3) -equation (4), we have the following functions:

$x_t^2 = \theta_t^2\ \varepsilon_t^2$

$\theta_t^2 = \alpha_0 + \alpha_1 x_t^2$.

For ARCH, if there is series, $\{ x_p, x_{p-1}, x_{p-2}, \ldots \}$, using equation (5), we obtain if the mean of $y_t$is zero:

E (x_t) = EE ($x_t \mid x_{t-1}$) = 0 (6)

Thus, for a martingale difference of x_t, there is the function:

Cov (x_{t+h}, y_x) = E ($x_{t+h} x$) = E $\{x_t$ E $x_{t+h} \mid x_{t+h-1}\}$ =0 (7)

For GARCH (1,1) model, it has the expression below:

$x_t^2 = \alpha_0 + (\alpha_1 + \beta_1) x_{t-1}^2 + v_t - \beta_1 v_{t-1}$, (8)

where x_t^2 is stationary for $\alpha_1 + \beta_1 < 1$.

C. Forecasting

To compute E($E(x_t^2)$, we can use the functions:

$x_t^2 = \theta_t^2 + Z_t = \omega + \sum_{j=1}^{q} \gamma_j Z_{t-j} + \sum_{i=1}^{R} (\theta_i + \gamma_j) x_{t-i}^2 + Z_t,$ (9)

where R=max (p, q), θ_i =0 for i > p, and γ_j=0 for j> q.

So, $x_{t+h}^2 = \omega + \sum_{i=1}^{R} (\theta_i + \gamma_j) x_{t+h-i}^2 - \sum_{j=1}^{q} \gamma_j Z_{t-j} + Z_{t+h}$, if E $(Z_{t+h} | F_k) = 0$

E $(x_{t+h}^2 | F_k) = \omega + \sum_{i=1}^{R} (\theta_i + \gamma_j) E(x_{t+h-i}^2 | F_k) - \sum_{j=1}^{q} \gamma_j E(Z_{t+h-j} | F_k) + E(Z_{t+h} | F_k)$

E $(x_{t+h}^2 | F_k) = \omega + \sum_{i=1}^{R} (\theta_i + \gamma_j) E(x_{t+h-i}^2 | F_k) - \sum_{j=1}^{q} \gamma_j E(Z_{t+h-j} | F_k)$ (10)

We often use equation (9) to compute the GARCH forecast, but it exists in some boundary conditions:

a. If h >*i*in equation (5), E $(y_{t+h}^2 | F_k)$ is given recursive by (10),
b. $E(x_{t+h-i}^2 | F_k) = x_{t+h-i}^2$, while h≤ j,
c. If h≤ j, $E(Z_{t+h-j} | F_k) = Z_{t+h-j}$,

If h >j, $E(Z_{t+h-j} | F_k)$=0.

2.2 ARIMA Forecasting Model

If a time series is y_{m+n}, m=1,2, …,y= { $y_n, y_{n-1}, \dots, y_1$} , ARIMA forecasting model has the function to predict future values of y_{m+n}:

$y_{m+n}^n = E\ (y_{m+n} | y)$, (11)

where y_{m+n}^n is the minimum mean square error predictor.

For one-step-ahead prediction, { $y_1, y_2, \dots, y_n$}, it has the function to forecast the value of the time series at the next time point, y_{n+1}:

$y_{n+1}^n = \Phi_n y_n + \Phi_{n2} y_{n-1} + \cdots + \Phi_{nn} y_1,$ (12)

2.3 Seasonal ARIMA Model

For ARMA $(P, Q)_s$, we have the form:

$\Phi_P(B^s) y_t = \Theta_Q(B^s) w_t,$ (13)

Then we get the definition:

$\Phi_P(B^s) = 1 - \Phi_1(B^s) - \Phi_2(B^s) - \cdots - \Phi_P(B^{Ps})$ (14)

and

$\Theta_Q(B^s) = 1 + \Theta_1(B^s) + \Theta_2(B^s) + \cdots + \Theta_Q(B^{Qs}),$ (15) where ARMA $(P, Q)_s$ is causal only when the roots of $\Phi_P(B^s)$ lies outside the unit circle and it is invertible only when the roots of $\Theta_Q(B^s)$ lies outside the unit circle, s is seasonal periods.

Hence, we can obtain the first-order seasonal ARIMA series that might be:

$y_t = \Phi y_{t-12} + w_t + \Theta w_{t-12}.$ (16)

The above equation displays the series y_t in terms of past lags at the multiple of the yearly seasonal periods s=12 months. Sometimes, the causal conditional defines $|\Phi| < 1$, but $|\theta| < 1$ is defined for the invertible condition.

2.4 Remarks on Methodology

ARIMA, SARIMA or GARCH model in time series method has been applied extensively in many research or journal papers. However, compressively studied for hybrid, such as two or more techniques ARIMA and SARIMA, or ARIMA-GARCH and SARIMA were rarely seen in journal papers or teaching. Hence, the author would like to explore to use them. For example, in figure 2, the author uses GARCH predictions by volatility method to compute and plot 1576 observations to analyze HIV time series data, so that obtain GARCH predictions of the HIV volatility, ± 2 σˆ_t, indicated as dashed lines with blue color; in Figure 5 and Figure 6, the author also applies ARIMA (0,0,0) and ARIMA (0,0,2) to forecast the future values after 2015, so we obtain good effects. These have confirmed that we could explore new methods or more means to deal with the complicated time series data.

3. Empirical Results

For any time series analysis of data, finding the best model that fits data is first thing. So, the author uses some steps: building a time plot of the data and check the graphs for any anomalies. Inspecting autocorrelation function (ACF) and partial ACF (PACF) are very important. The ACF function is:

$$\rho(s, t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}.$$

Hence, the ACF of stationary time series is: $\rho(\quad) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}}$.

The ACF is used to assess the linear predictability of the series at time t, y_t, which is used only the value y_s, here $-1 \leq \rho(s,t) \leq 1$ for all h(the Cauchy-Schwarz inequality). Also, since there is a linear relationship, $y_t = \beta_0 + \beta_1 y_s$, we might predict y_t from y_s if the correlation is +1 for $\beta_1 > 0$, and -1 for $\beta_1 < 0$. Thus, at that time we might assess the ability to predict the series at time t from the value at time s. In this case, inspecting the ACFand PACFof HIV data and residuals: they show to be abnormal.

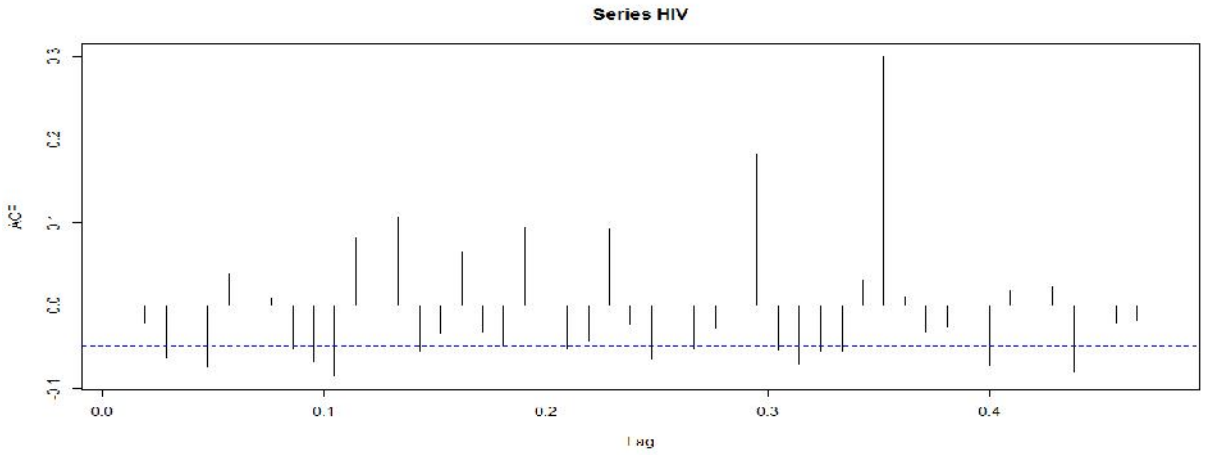

Figure 1. ACF for first 50 lags. It looks like that they are over base line 0.1.

On the other hand, in inspect the model, we need to check Akaike's Information Criterion (AIC). AIC =log $\hat{\sigma}_k^2 + \frac{n+2k}{n}$, where $\hat{\sigma}_k^2$is given by $\hat{\sigma}_k^2 = \frac{SSE_k}{n}$ in whichSSE_k expresses the residual sum of squares under the model with k of the number of parameters in the model.Using the R package f GARCH is to fit an GARCH (1,0) model for HIV returns. A partial output is as follows:

```
Title:
 GARCH Modelling

Call:
 garchFit(formula = ~arma(2, 0) + garch(1, 0), data = HIV.data)

Mean and Variance Equation:
 data ~ arma(2, 0) + garch(1, 0)
<environment: 0x102751a8>
 [data = HIV.data]

Conditional Distribution:
 norm

Coefficient(s):
        mu         ar1         ar2       omega      alpha1
 0.0081884  -0.6918606  -0.2636078   4.3298039   0.0708400

Std. Errors:
 based on Hessian

Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
mu      0.008188    0.055679    0.147    0.883
ar1    -0.691861    0.027695  -24.981   <2e-16 ***
ar2    -0.263608    0.023652  -11.145   <2e-16 ***
omega   4.329804    0.242901   17.825   <2e-16 ***
alpha1  0.070840    0.044959    1.576    0.115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Standardised Residuals Tests:
                                Statistic p-Value
 Jarque-Bera Test   R    Chi^2  6.9972    0.03024
 Shapiro-Wilk Test  R    W      0.99029   1.0001e-08
 Ljung-Box Test     R    Q(10)  342.05    0
 Ljung-Box Test     R    Q(15)  472.71    0
 Ljung-Box Test     R    Q(20)  494.28    0
 Ljung-Box Test     R^2  Q(10)  180.32    0
 Ljung-Box Test     R^2  Q(15)  203.74    0
 Ljung-Box Test     R^2  Q(20)  251.18    0
 LM Arch Test       R    TR^2   208.94    0

Information Criterion Statistics:
   AIC    BIC    SIC   HQIC
4.3791 4.3961 4.3791 4.3854
```

The results indicate that $\hat{\phi}_0 = -0.0082$ (mu), $\hat{\phi}_1 = -0.692$ (ar1) for AR (2) parameter estimate. The ARCH (1) parameter estimates are $\hat{\alpha}_0$=4.330 (omega) for the constraint and $\hat{\alpha}_1$ =0.071, which is significant with a p-value of about 0.05. For Ljung-Box tests, all of p-values are less than 0.05. These specifications show that ARIMA (2, 0)- GARCH (1,0) is the best model to analyze and predict the HIV data.

However, in this article, there are some of tests performed by the residuals and /or the squared residuals. Since the skewness and kurtosis are required to observe, some residuals suitable for normality would be tested by the JarqueBera. Also, some tests such as the Shapiro-Wilk should test those residuals based on the empirical order statistics. Some tests, which are the residuals and their squires, need to check the Q-statistic.

The above R output of GARCH (1,1) suggested that “ar 1”, “ar 2”, and “Omega” have significant values ($p < 0.05$).

```
Title:
 GARCH Modelling

Call:
 garchFit(formula = ~garch(1, 1), data = HIV)

Mean and Variance Equation:
 data ~ garch(1, 1)
<environment: 0x09b70f88>
 [data = HIV]

Conditional Distribution:
 norm

Coefficient(s):
        mu       omega      alpha1       beta1
2.4094e+01  1.8739e-01  1.0000e-08  1.0000e+00

Std. Errors:
 based on Hessian

Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
mu     2.409e+01          NA       NA       NA
omega  1.874e-01   4.255e+00 4.40e-02    0.965
alpha1 1.000e-08   8.736e-04 0.00e+00    1.000
beta1  1.000e+00   3.356e-05 2.98e+04   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Log Likelihood:
 -8918.4    normalized:  -5.6589
```

The output of GARCH (1,1) displays that “beta1 is less than 0.05. But, in the following one, it indicates that some important test values are less than 0.05, such as Jarque-Bera test, Shapiro-Wilk test, etc.

```
Standardised Residuals Tests:
                                Statistic p-Value
 Jarque-Bera Test   R    Chi^2  207574    0
 Shapiro-Wilk Test  R    W      0.33682   0
 Ljung-Box Test     R    Q(10)  32.709    0.00030485
 Ljung-Box Test     R    Q(15)  77.941    1.6585e-10
 Ljung-Box Test     R    Q(20)  105.87    1.1069e-13
 Ljung-Box Test     R^2  Q(10)  3.1879    0.97665
 Ljung-Box Test     R^2  Q(15)  4.8114    0.99361
 Ljung-Box Test     R^2  Q(20)  6.1042    0.99875
 LM Arch Test       R    TR^2   4.1197    0.98118

Information Criterion Statistics:
   AIC    BIC    SIC   HQIC
11.323 11.337 11.323 11.328
```

To analyze the GARCH predictions by volatility, the author calculated and plotted 1576 observations from HIV time series data. For the one-step-ahead predictions for volatility, σ_t^2, the results showed as the data $\pm$ 2 $\hat{\sigma}_t$by dashed linessurrounding the data in Figure 3.Nevertheless, some of the GARCH models include: (a) since volatility is based on squared returns, the model assumes the returns from positive and negative are possible to get the same effect. (b) It is flat for the likelihood unless n is very large. (c) If it is large isolated return, model responds slowly, so the model is overpredict volatility. (d) Sometimes since it is the tight constraints on the parameters, it perhaps to appear restrictive model, for example, an ARCH (1), $0 \leq \alpha_1^2 < \frac{1}{3}$.

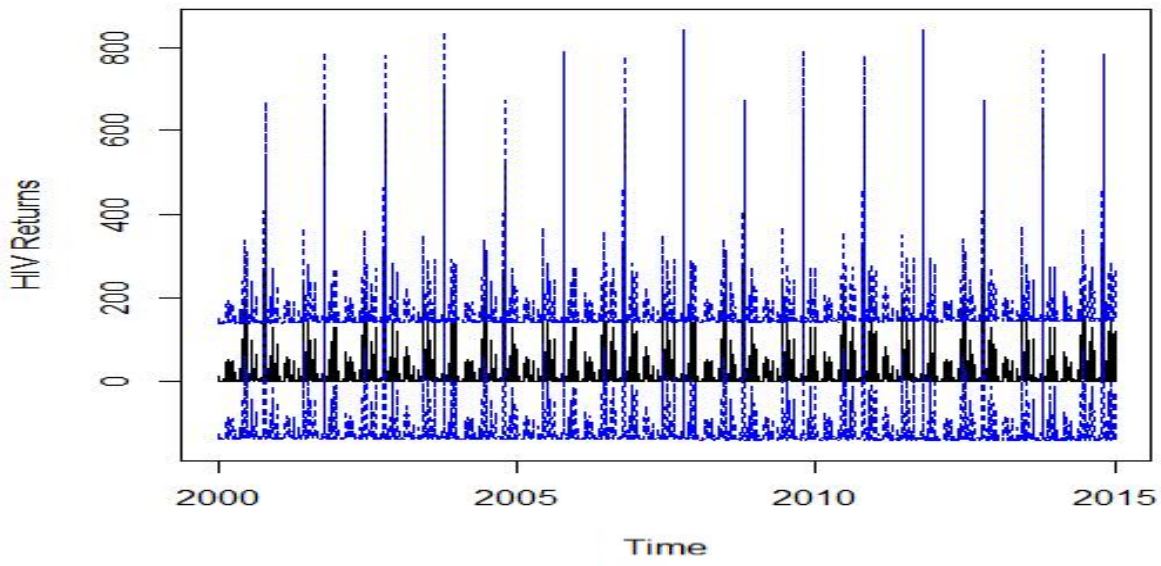

Figure 2. GARCH predictions of the HIV volatility, $\pm$ 2 $\hat{\sigma}_t$, indicated as dashed lines with blue color.

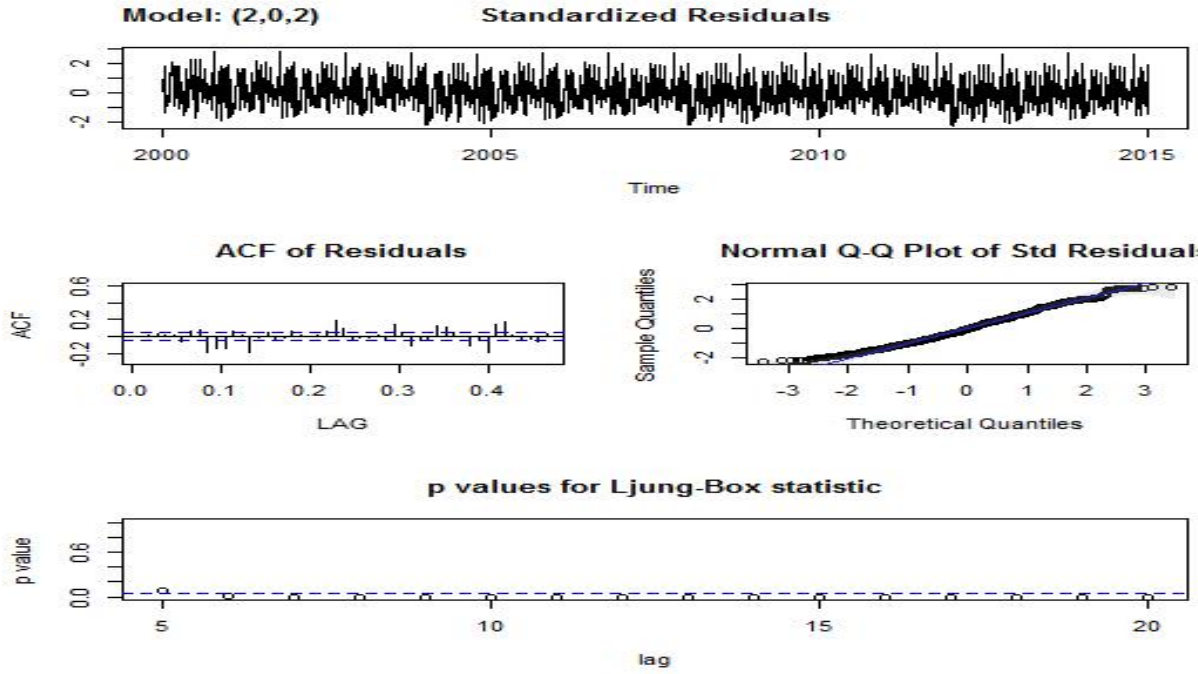

Figure 3. Q-statistic p-values for the ARIMA (2, 0, 2) fit to the logged HIV data, QQ plot from Residuals is good.

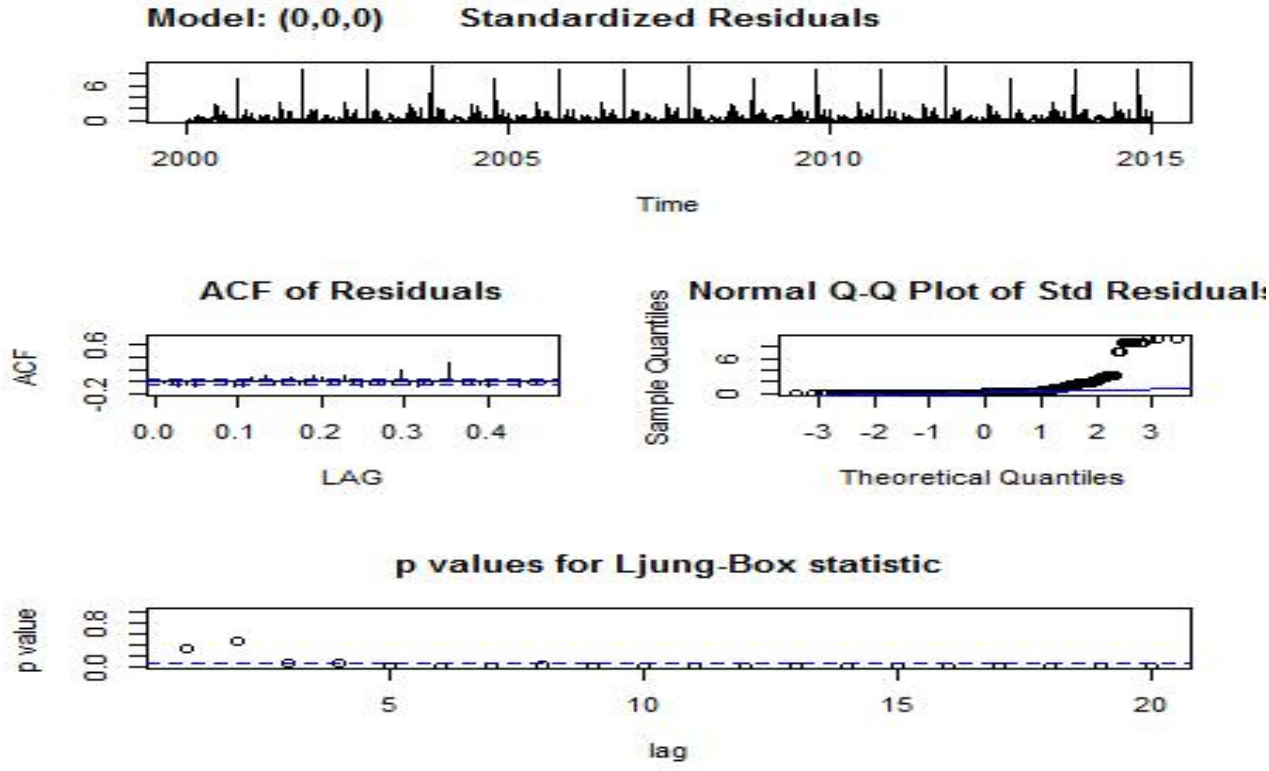

Figure 4. SARIMA plot of original HIV data. Diagnostics for ARIMA (0, 0, 0).

If we take look the following figures (Figure 5 and Figure 6), it is not hard to see that they have difference in forecasting the values and trends between both using ARIMA(0, 0, 0) and Seasonal ARIMA (0, 0, 0).

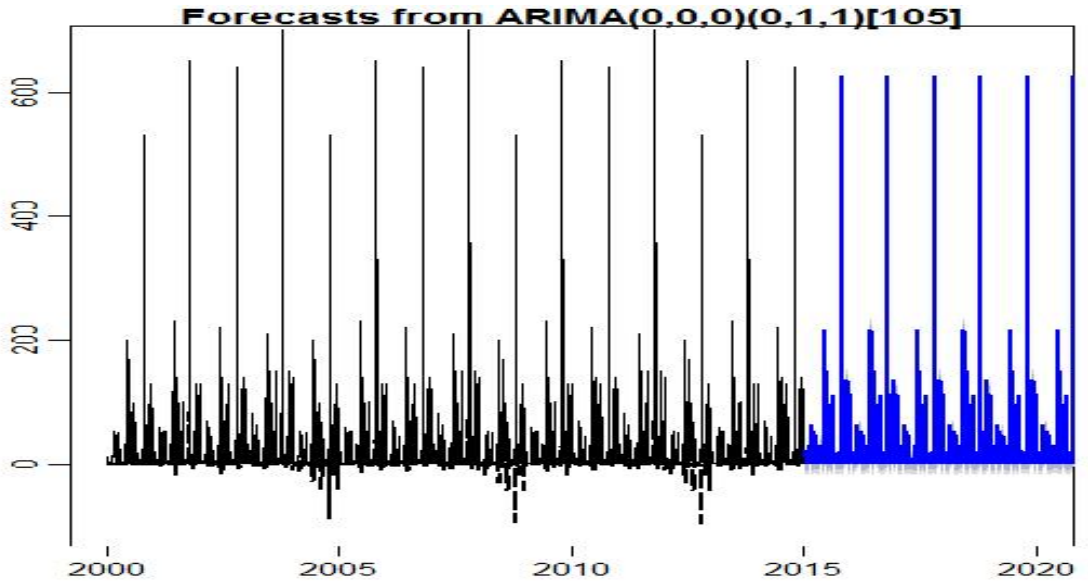

Figure 5. Forecasting plot of HIV time series data using ARIMA (0,0,0). It exhibits that the future values and trends looks alike to the value before 2015.

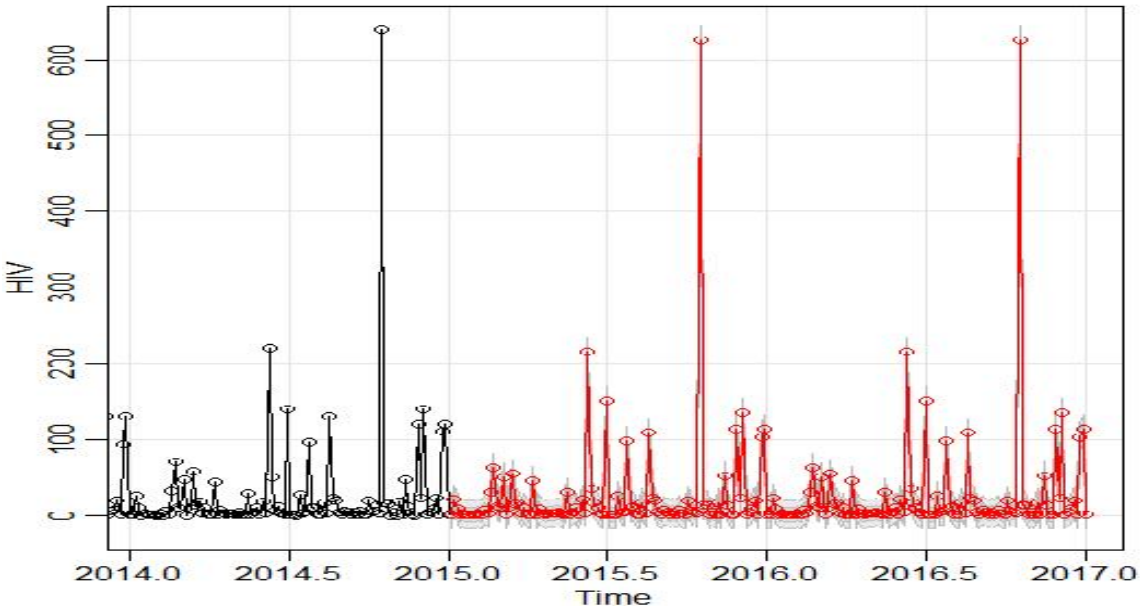

Figure 6. Forecasting plot of original HIV datausing seasonal ARIMA (0, 0, 2). It displays the results of forecasting future values resembles with forecasting using ARIMA (0, 0, 0).

4. Conclusion

ARIMA-GARCH mode is one of most popular statistical methods to analyze and predict time series data. It is very useful to various fields and research. ARIMA model is well-knownand simpler statistical methods. Seasonal ARIMA method can be used to compare with other statistical models such other methods as ARIMA.When a model is hard to be estimated accurately. We might to try used another one or more to compare the difference in analysis and prediction.In this case, the author applies auto.arima function to determine the best model, that is, ARIMA(0, 0, 0), and then let function garchFit (~2,0)+garch (1,0) analyze the HIV data, finally, uses both ARIMA predict functionand seasonal ARIMA. for predict model to obtain same forecasting results. The author thinks that using ARIMA-GARCH model could be the best way to analyze time series data and then to utilize some other models such as ARIMA, seasonal ARIMA, Holt-Winters, other methods to compare the results of forecasting the future values and trends. This is a very helpful and valuable work.

Acknowledgements

The Author thanks editor and two anonymous reviewers. They gave constructive suggestions and helps.

Competing Interest

Not the author declares competing interests.

References

1. World Health Statistics data visualizations. World Health Organization 2015, Geneva, Switzerland.
2. Robert H. Shumway, David S. Stoffer. Time Series analysis and its application with R examples. 3rd ed. New York, NY: Springer Science; 2011.
3. Liming Xie. Time Series Analysis and Prediction on Cancer Incidence Rates. Journal of Medical Discovery, 2017 2(3):jmd17030; DOI:10.24262/jmd.2.3.17030.
4. Robert H. Shumway, David S. Stoffer. Time Series analysis and its application with R examples. 3rd ed. New York, NY: Springer Science; 2011.
5. Peter J. Brockwell. and Richard. A. Davis. Introduction to Time Series and Forecasting. 2nd ed. New York, NY: Springer; 2002.
6. Robert H. Shumway, David S. Stoffer. Time Series analysis and its application with R examples 3rd ed. New York, NY: Springer Science.; 2011: 281 -282.
7. Robert H. Shumway, David S. Stoffer. Time Series analysis and its application with R examples 3rd ed. New York, NY: Springer Science; 2011: 154 -155.
8. Zhongfu Tan, Jinliang Zhang, Jianhui Wang, et al.Day-ahead electricity price forecasting using wavelet transform combined with ARIMA and GARCH models. Applied Energy. 2010; 87:3606-3610.
9. Mohamadi, Salman, Hamidreza, et al. ARIMA-GARCH modeling for epileptic seizure prediction. IEEE International Conference on Acoustics, Speech and Signal Process. 2017; 994-998.

10. Li Ming-Jia, He YapLing, Tao Wen-Quan. Modeling a hybrid methodology for evaluating and forecasting regional energy efficiency in China. Applied Energy. 2017;185: 1769-1777.
11. Amiri, Esmail. Empirical study of GARCH models with leverage effect in an environmental application. Environmental and Ecological Statistics. 2014;21:125-141.
12. Hassan Mohammadi, Lixian Su. International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models. Energy Economics. 2010; 32:1001-1008.
13. KamilMakiel. ARIMA-GARCH models in estimating market risk using value at risk for the WIG 20 index. E-Finanse. 2012; 8:25-33.
14. Gordon W, Crawford, Michael C. Fratantoni. Assessing the forecasting performance of Regime-Switching, ARIMA and GARCH models of house prices. Real Estate Economics. 2003; 31: 223-243.
15. Auer, Benjamin R. Superstitious seasonality in precious metals markets? Evidence from GARCH models with time-varying skewness and kurtosis. Applied Economics. 2005;2: 1-16.
16. Mohan Kumar, T.L, Prajneshu, T.L. Development of hybrid models for forecasting time series using nonlinear SVR enhanced by PSO. Journal of Statistical Theory and Practice.2015; 7:1-13.
17. Ruey S. Tsay. An Introduction to Analysis of Financial Data with R. New Jersey, Jon Wiley &Sons. Inc; 2013:199-200.
18. Soren Bisgaard, Murat Kulahci. Time Series Analysis and Forecasting by Example. New Jersey, Jon Wiley &Sons. Inc; 2011:162 -163.
19. Nicholas J. Horton, Ken Kleinman. Using R and R studio for Data Management, Statistical Analysis, and Graphics.2nd ed. Boca Raton, FL; 2015: 132 -133.