Scientific
Research
Publishing

# Using Box-Jenkins Models to Forecast Mobile Cellular Subscription

## Ian Siluyele[1], Stanley Jere[2]

[1]Zambia Information and Communications Technology Authority (ZICTA), Lusaka, Zambia
[2]Department of Mathematics and Statistics, Mulungushi University, Kabwe, Zambia
Email: isiluyele@zicta.zm, sjere@mu.ac.zm

## Abstract

In this paper, the Box-Jenkins modelling procedure is used to determine an ARIMA model and go further to forecasting. The mobile cellular subscription data for the study were taken from the administrative data submitted to the Zambia Information and Communications Technology Authority (ZICTA) as quarterly returns by all three mobile network operators Airtel Zambia, MTN Zambia and Zamtel. The time series of annual figures for mobile cellular subscription for all mobile network operators is from 2000 to 2014 and has a total of 15 observations. Results show that the ARIMA (1, 2, 1) is an adequate model which best fits the mobile cellular subscription time series and is therefore suitable for forecasting subscription. The model predicts a gradual rise in mobile cellular subscription in the next 5 years, culminating to about 9.0% cumulative increase in 2019.

## 1. Introduction

In Zambia, the penetration of information and communication technology (ICT) in general and mobile in particularly, plays an important role in compilation of the national Gross Domestic Product (GDP). There are three (3) mobile cellular operators in Zambia with networks spanning land area of almost 602,090 $Km^2$, representing 80% network coverage. In 2014, Zambia had 67.1% of subscribers from the mobile cellular subsector with revenue contribution of nearly K3.4 billion. At the end of December 2014 the population of Zambia was estimated at 15.1 million while mobile cellular subscription (MCS) was 10.1 million.

Studies have shown that diffusion of mobile telecommunication affects the growth of GDP. Other studies have also shown that a long run causal relationship exists between growth in telecommunications and the growth

of the economy both at sectoral and aggregate levels. Therefore, the importance of investment in telecommunication subsector is acknowledged world over. Globally, socio-economic effect and economic development due to improved telecommunication cannot be repudiated.

Time series modelling is an important part of every field. It provides both short and long term forecasting techniques. Effective implementation of forecasting techniques maximises the prospect of adopting optimum strategies. Literature shows that researchers have used both stochastic and deterministic models to model and forecast telecommunication data. However, stochastic models attributed to Box-Jenkins, the Auto Regressive Integrated Moving Average (ARIMA) models have been found to be more efficient and reliable even for short term forecasting than the deterministic models. Further, stochastic models are distribution-free as no assumptions are required about the data or parameter hence the adoption of the forecasting methodology in this paper.

## 2. Method and Materials

The MCS data for the study has been taken from the administrative data submitted to the Zambia Information and Communications Technology Authority (ZICTA) as quarterly returns by all three mobile network operators (MNOs). The time series of annual figures for MCS for all MNOs is from 2000 to 2014 and has a total of 15 observations. Each observation ($X_t$) in the time series is sum total of subscriber for Airtel Zambia, MTN Zambia and Zamtel *i.e.*

$$X_t = \alpha_t + \beta_t + \gamma_t t = 2000, \cdots, 2014$$

where,

$\alpha_t$ = number of subscribers for *Airtel Zambia*,
$\beta_t$ = number of subscribers for *MTN Zambia*, and
$\gamma_t$ = number of subscribers for *Zamtel*.

Statistical Analysis System (SAS) [1] and Microsoft Excel will be used to implement the stochastic models and graphical representations respectively.

## 3. Stochastic Modelling

The Box-Jenkins approach to forecasting was first described by statisticians George Box and Gwilym Jenkins and was developed as a direct result of their experience with forecast problems in the business, economic, and control engineering applications [2]. The Box-Jenkins methodology is a systematic process which is implemented by using an iterative process until an adequate model is achieved. The procedure is achieved by a step-by-step process of model IDENTIFICATION, SPECIFICATION, ESTIMATION, DIAGNOSTIC and FORECAST. The ARIMA has three parameters viz. autoregressive ($p$), differencing order ($d$) and the order of moving average ($q$). The generic Box-Jenkin models are denoted by ARIMA ($p$, $d$, $q$) given by

$$\varnothing(B)(X_t - \mu) = \theta(B)a_t$$

where, $\varnothing$, $\theta$ and $a$ are autoregressive parameter, moving average parameter and residual respectively. The residuals are assumed to be iid[1] Normal. Using the backshift operator/transformation the equation above, when $d = 0$, is expressed as

$$\left(1 - \varnothing_1 B^1 - \cdots - \varnothing_p B^p\right) X_t = \left(1 - \varnothing_1 B^1 - \cdots - \varnothing_p B^p\right) \mu + \left(1 - \theta_1 B^1 - \cdots - \theta_q B^q\right) a_t.$$

Also as

$$X_t = \text{Constant} + \frac{\theta(B)a_t}{\varnothing(B)}$$

where, $a_t$ is a white noise process with mean 0 and variance $\sigma^2$ [3].

## 4. Measures of Forecast Accuracy

The statistics are used to compare how well models fit the time series. Akaike Information Criterion (AIC) and

---

[1]Independent and identically distributed.

Schwartz's Bayesian Information Criterion (SBC) are some of the measures of accuracy of forecast that are widely used in SAS. Other measures used include Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Deviation (MAD). Forecast error is given by

$$\text{Forecast Error}\left(FE_t\right) = \text{Observed Value}\left(O_t\right) - \text{Forecast Value}\left(F_t\right).$$

AIC and SBC are given by

$$\text{AIC} = -2\left(\ln\left(L\right) - k\right) \quad \text{and} \quad \text{SBC} = -2\ln\left(L\right) + k\ln\left(n\right).$$

Other measures of forecast accuracy are given by

$$\text{MSE} = \frac{\sum_n^1 \left(O_t - F_t\right)^2}{n}, \quad \text{MAPE} = \frac{\sum_n^1 \left|\frac{FE_t}{O_t}\right|}{n} \times 100 \quad \text{and} \quad \text{MAD} = \frac{\sum \left|FE_t\right|}{n}.$$

In these formulas, $L$ is the value of the likelihood function evaluated at the parameter estimates, $n$ is the number of observations, $k$ is the number of estimated parameters and $t = 1, 2, \cdots, n$.

## 5. Identification

This is the foremost step of the Box-Jenkins process of time series modelling. A timeplot of the MCS is plotted in **Figure 1(a)** and checked for stationarity and invertibility using visual display of the ACF[2] and PACF[3] graphs.

**Figure 1(a)** show that the MCS time series is not stable and therefore nonstationary. The nonstationary behaviour is confirmed by the ACF and PACF plots in **Figure 2(a)** and **Figure 2(b)** below. Therefore some sort of transformation of the series is necessary to make it mean and variance stationary[4]. ARIMA models are designed to model stationary time series.

Converting a nonstationary time series to a stationary one through differencing (where needed) is an important part of the process of fitting an ARIMA model. **Figure 1(b)** and **Figure 1(c)** shows first order and second order differenced MCS series, respectively. The ACF and PACF plots are shown in **Figure 2(c)**, **Figure 2(d)**, **Figure 2(e)** and **Figure 2(f)** below. ACF and PACF plots indicate that the first and second differenced MCS series are stationary hence require further examination to establish the most suitable transformation for the MCS series.

**Table 1** shows the details of various ARIMA models along the forecast accuracy measures. An ARIMA model with least measures of accuracy particularly the AIC and SBC is considered an efficient model for prediction. Therefore, for MCS time series, the **ARIMA (1, 2, 1)** is an adequate (best fit) model because it has the lowest values for AIC and SBC statistics.

## 6. Parameter Estimation

**Table 2** shows the estimated parameters and the associated p-values at 5% level of significance. Only the autoregressive parameter is significantly different from zero at 5% implying that the constant and the parameter for the moving average coefficients have little or no effect on the model.

The model variable and factors are given in **Table 2**. Hence, the mathematical form of the **ARIMA (1, 2, 1)** is

$$\left(1 - B\right)^2 X_t = \frac{a_t}{\left(1 - 0.54647B\right)}.$$

## 7. Diagnistic Check

Verification of goodness of fit of any model should include a test as to whether the residuals form a white noise process. Diagnostic check helps determine if an estimated model is statistically adequate. If the identified model passes the diagnostic tests, the model is ready to be used for forecasting. If it does not, the diagnostic tests

---

[2]Autocorrelation function.
[3]Partial autocorrelation function.
[4]A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time.
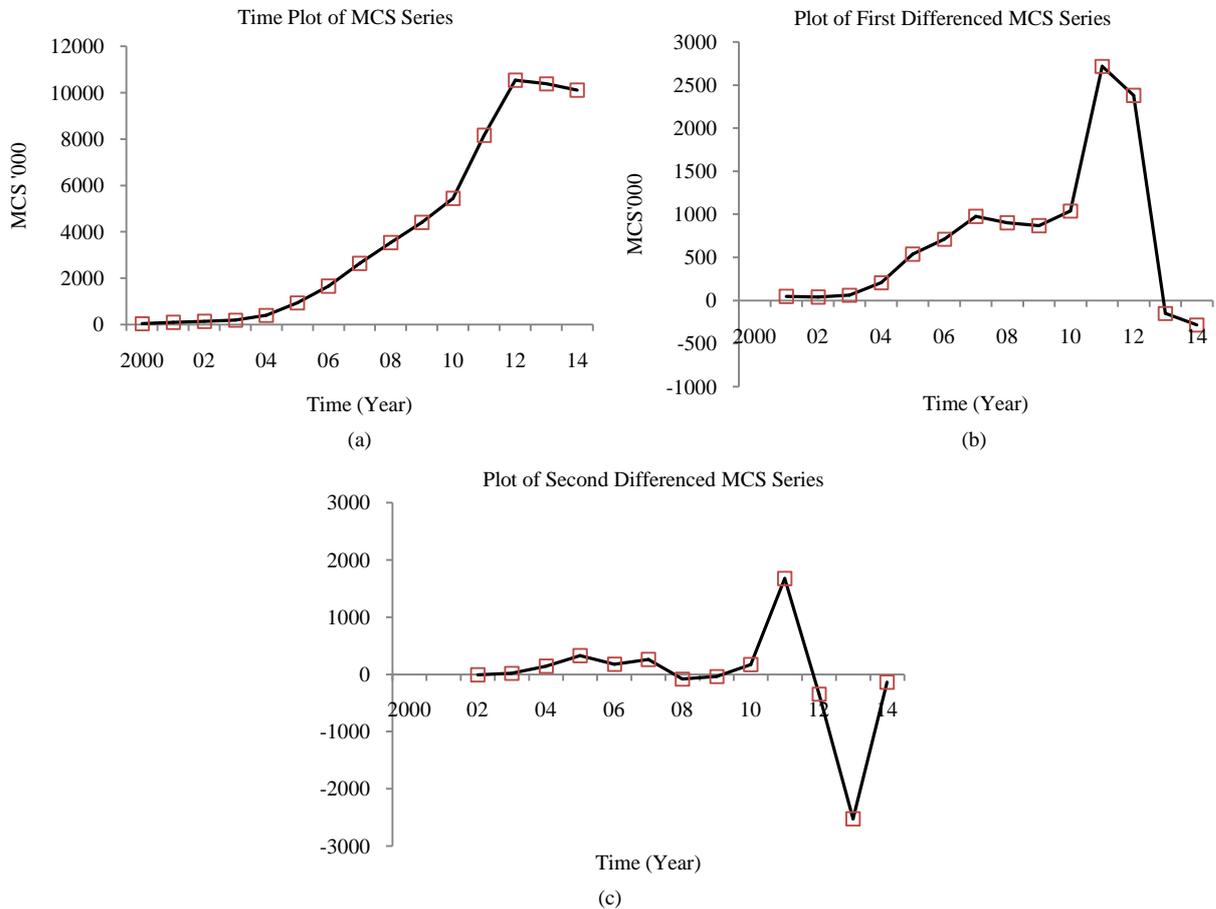
Figure 1. Time plots for d = 0, d = 1 and d = 2 of MCS series.

Table 1. Measures of accuracy for selected ARIMA models.

| ARIMA MODELS | AIC | SBC | RMSE | MAPE | MAD |
|---|---|---|---|---|---|
| ARIMA (1, 1, 0) | 422.68 | 423.96 | 725,927.03 | 79.94 | 500,531.81 |
| ARIMA (1, 1, 1) | 423.06 | 424.98 | 682,365.52 | 89.63 | 444,145.63 |
| ARIMA (1, 1, 2) | 425.07 | 427.63 | 679,685.36 | 87.49 | 436,179.08 |
| ARIMA (2, 1, 1) | 425.00 | 427.56 | 681,695.68 | 95.37 | 448,809.15 |
| ARIMA (1, 2, 0) | 404.26 | 405.39 | 1,086,694.05 | 93.33 | 772,506.46 |
| **ARIMA (1, 2, 1)** | **399.91** | **401.61** | **783,904.98** | **46.22** | **525,306.06** |
| ARIMA (2, 2, 1) | 400.35 | 402.61 | 770,165.27 | 87.90 | 530,058.84 |
| ARIMA (2, 2, 2) | 406.07 | 408.89 | 937,548.54 | 116.59 | 684,186.40 |

Table 2. Estimated parameter and significance tests.

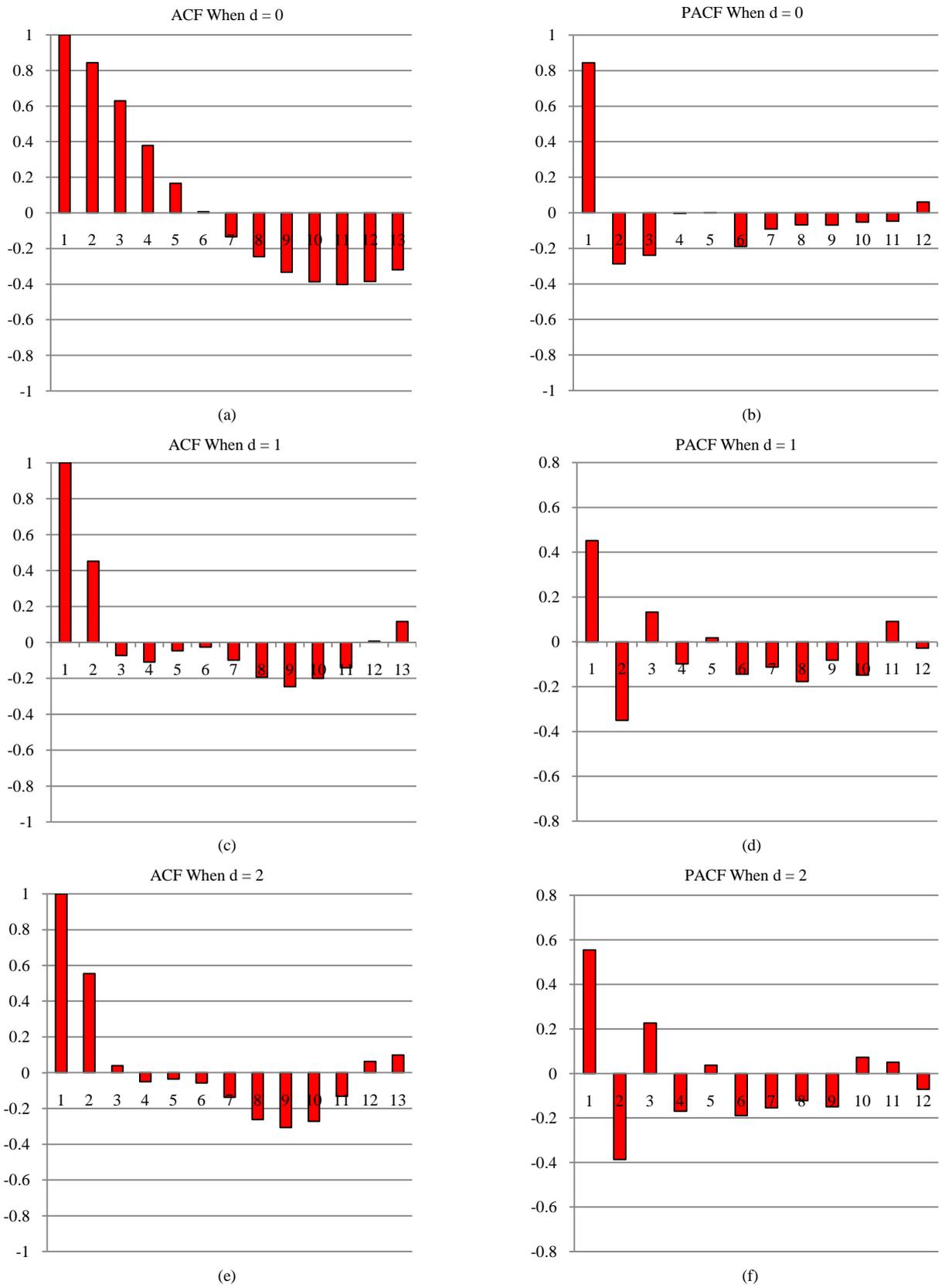| TYPE | Coefficient | SE coefficient | t-statistics | p-value | Lag |
|---|---|---|---|---|---|
| Constant | 500,796 | 967,576 | 0.52 | 0.6048 | 0 |
| MA (1) | −0.99995 | 111.0529 | −0.01 | 0.9928 | 1 |
| AR (1) | 0.54647 | 0.2761 | 1.98 | 0.0477 | 1 |

**Figure 2.** The ACF and PACF plots for d = 0, d = 1 and d = 2 of MCS series.

should indicate how the model ought to be modified, and a new cycle of identification, estimation and diagnosis is performed.

The Autocorrelation check for white noise of an **ARIMA (1, 2, 1)** model in **Table 4** p-values at 5% level of significance as shown above indicates that the model is good because the residuals are a white noise.

## 8. Forecasting

Box-Jenkins approach to forecasting stationary time series is relatively simple. The forecast value of $X_{t+k}$ given all observations up until $n$ the $k$-step ahead forecast is denoted by $\hat{x}_t(k)$. **Table 3** shows five year forecasts for mobile cellular subscription using **ARIMA (1, 2, 1)**. The trajectory of the forecasts from 2015 to 2019 is shown in **Figure 3**.

**Table 3.** Autocorrelation check for white noise.

| To lag | Chi-square | DF | Pr > ChiSq | Autocorrelations | | | | | |
|--------|-----------|-----|-----------|--------|--------|--------|--------|--------|--------|
| 6 | 5.72 | 6 | 0.4558 | 0.555 | 0.040 | −0.051 | −0.035 | −0.057 | −0.139 |
| 12 | 18.63 | 12 | 0.0979 | −0.263 | −0.307 | −0.272 | −0.132 | 0.063 | 0.098 |

**Table 4.** Forecasts for **ARIMA (1, 2, 1)** for mobile cellular subscription.

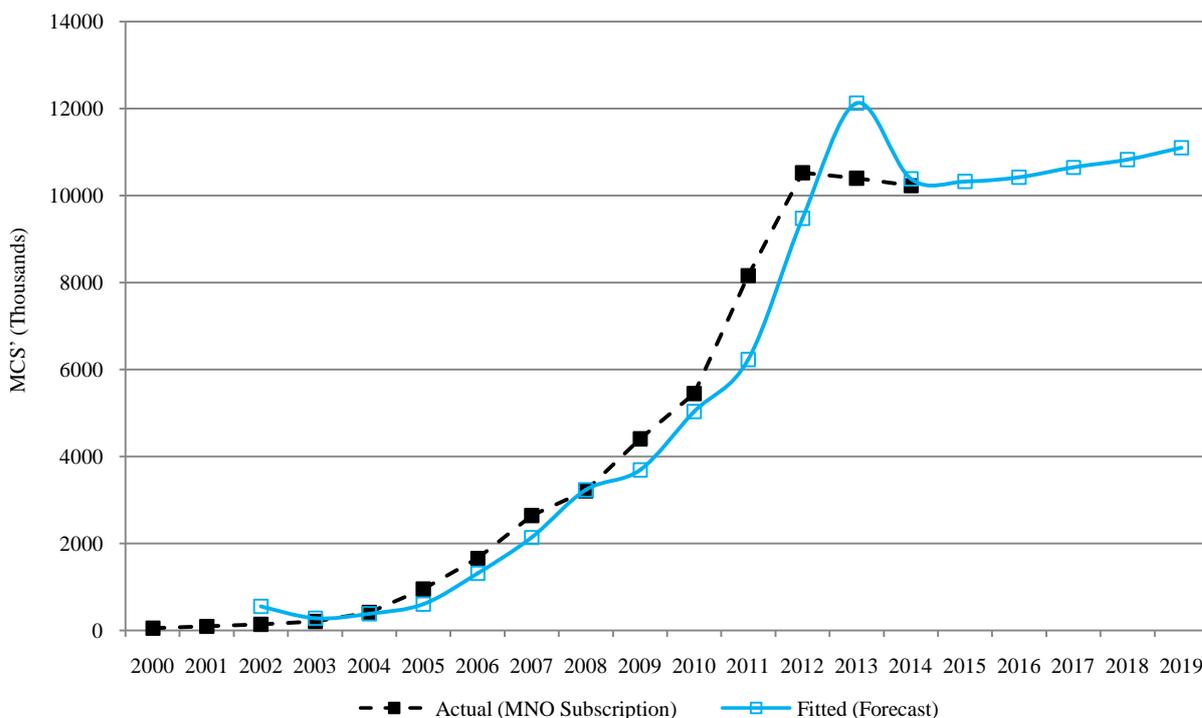| Forecasts for mobile cellular subscription | | | | |
|---|---|---|---|---|
| Year | Forecast | Std error | 95% confidence limits | |
| 2015 | 10,321,672 | 905,295 | 8,547,326 | 12,096,019 |
| 2016 | 10,418,227 | 1,667,167 | 7,150,640 | 13,685,815 |
| 2017 | 10,650,781 | 2,359,965 | 6,025,333 | 15,276,228 |
| 2018 | 10,825,202 | 2,979,034 | 4,986,402 | 16,664,002 |
| 2019 | 11,100,306 | 3532,708 | 4,176,325 | 18,024,287 |



**Figure 3.** Forecasting with **ARIMA (1, 2, 1)** for mobile cellular subscription.

## 9. Discussion

The **ARIMA (1, 2, 1)** is an adequate model which best fits the mobile cellular subscription time series and is therefore suitable for forecasting subscription. The potential implication of this study is that by developing forecasting models for predicting mobile cellular subscription in advance on a regular basis is to support internal decisions and planning as well as market communication. The subscription forecast baseline in this study uses historical data from Airtel Zambia, MTN Zambia and Zamtel. The study also provides a model to foresee and allocate appropriate resources to maintain a steady increase in mobile cellular subscription.

## 10. Conclusion

In this paper, the Box-Jenkins modelling procedure is used to determine an ARIMA model and go further to forecasting. The mobile cellular subscription data for the study were taken from the administrative data submitted to the Zambia Information and Communications Technology Authority (ZICTA) as quarterly returns by all three mobile network operators Airtel Zambia, MTN Zambia and Zamtel. The time series of annual figures for mobile cellular subscription for all mobile network operators is from 2000 to 2014 and has a total of 15 observations. Results show that the **ARIMA (1, 2, 1)** is an adequate model which best fits the mobile cellular subscription time series and is therefore suitable for forecasting subscription. The model predicts a gradual rise in mobile cellular subscription in the next 5 years, culminating to about 9.0% cumulative increase in 2019.

## Acknowledgements

## References

[1]  SAS Institute Inc. (2014) SAS/ETS 13.2 User's Guide: The ARIMA Procedure. SAS Institute Inc., Cary.

[2]  Box, G.E. and Jenkins, G.M. (1994) Time Series Analysis: Forecasting and Control. 3rd Edition, Prentice Hall, Englewood Cliffs.

[3]  Wei, W. (1990) Time Series Analysis: Univariate and Multivariate Methods. Addison-Wesley Publishing Company, Inc., New York.