Scientific Research Publishing

Modelling HIV/AIDS Cases in Zambia: A Comparative Study of the Impact of Mandatory HIV Testing

Edwin Moyo[1]*, James C. Shakalima[1], Gilbert Chambashi[2], James Muchinga[3], Levy K. Matindih[1]

[1]Department of Science and Mathematics, Mulungushi University, Kabwe, Zambia
[2]School of Business Studies, Unicaf University, Lusaka, Zambia
[3]Department of Natural Sciences, Levy Mwanawasa University, Lusaka, Zambia
Email: *emoyo@mu.ac.zm

Abstract

In this study, a time series modeling approach is used to determine an ARIMA model and advance counterfactual forecasting at a point of policy intervention. We consider monthly data of HIV/AIDS cases from the Ministry of Health (Copperbelt province) of Zambia, for the period 2010 to 2019 and have a total of 120 observations. Results indicate that ARIMA (1, 0, 0) is an adequate model which best fits the HIV/AIDS time series data and is, therefore, suitable for forecasting cases. The model predicts a reduction from an average of 3500 to 3177 representing 14.29% in HIV/AIDS cases from 2017 (year of policy activation) to 2019, but the actual recorded cases dropped from 3500 to 1514 accounting for 57.4% in the same time frame.

1. Introduction

Human Immunodeficiency Virus (HIV) infection remains the prominent cause of morbidity and mortality throughout the world. Since the start of the epidemic, around 76.1 million people have been infected and almost 35.4 million people have died from Acquired Immunodeficiency Syndrome (AIDs) related illnesses.

Globally, in 2019 there were an estimated 38 million people living with HIV. The vast majority of people living with HIV are located in low and middle-income countries, with an estimated 66% living in sub-Saharan Africa. Among this group

19.6 million are living in East and Southern Africa which saw increasing HIV infections due to HIV service disruptions during COVID-19 and the slowing public health response to HIV [1].

As one of the Sub-Saharan countries, the cases in Zambia are not different and HIV prevalence in Zambia continued to decline. The recent Zambia Population HIV Impact Assessment (ZAMPHIA) survey shows a reduction in the prevalence of about 1.7 percentage points from 13.3% in 2014 to 11.6% in 2016 [2]. The health sector has also recorded remarkable progress on antiretroviral treatment (ART) coverage, which stands at 72% of the eligible people against the United Nations AIDS global target of 90% according to Zambia National Health Strategic Plan (2017-2021) [2]. In 2018, around 48,000 adults and 5400 children became newly infected with HIV in Zambia. In the same year around 1.2 million people in Zambia were living with HIV and 17,000 people died from HIV-related illnesses. Also, the number of people newly diagnosed with HIV and with active tuberculosis (TB) entering care has fallen significantly from 66% in 2015 to 10% in 2017 with reference to [3].

Reference [3] articulates that in 2018, 78% of all people living with HIV were on treatment. As of 2019, 87% of people living with HIV were aware of their status, and 89% on treatment and 75% were virally suppressed. [3] also urges that in Zambia there has been progress in the number of AIDS-related deaths since 2010, with a 37% decrease, from 26,000 deaths to 17,000 deaths. The number of new HIV infections has also decreased, from 56,000 to 48,000 in the same period. The 90-90-90 targets envision that, by 2030, 90% of people living with HIV will know their HIV status, 90% of people who know their HIV-positive status will be accessing treatment and 90% of people on treatment will have suppressed viral loads [3].

The policy intervention on HIV/AIDS cases in Zambia presents a comparative (also known as counterfactual or causal impact analysis) time series problem. Reference [4] explains that this technique is used to study the causal effect resulting from the difference between the observed series and the series that would have been observed had the intervention not taken place. In this study, the Autoregressive Integrated Moving Average (ARIMA) model also known as the "Box-Jenkins" methodology following the work of Box and Jenkins was used to determine the ideal model of best fit to be used in forecasting the HIV/AIDS data. Box-Jenkins forecasting is of greatest use when the primary factors causing demand for products, services, revenue, and, in this instance, disease burden is believed to behave in the future as well as it did in the past [5].

The study by [6] evaluated the impact of cigarette exercise tax increases in Pennsylvania, from 2000 to 2009. They used segmented regression analysis of interrupted time series to evaluate the level and trend changes in Pennsylvania adult's current smoking prevalence, age-adjusted acute myocardial infarction (AMI) and asthma hospitalization rates, age-specific asthma hospitalization, and age adjusted rates following the cigarette exercise tax increases. Results showed

that after the first tax increase, no beneficial effects were noted on the outcome of interest. The second tax increase was associated with significant declines in smoking prevalence for people aged 18 to 39. A decrease was also noticed in asthma hospitalizations and men's aged-adjusted AMI hospitalization rates.

[7] employed intervention (comparative study) time series analysis to evaluate the effectiveness of a collaborative intervention to improve quality in pre-hospital ambulance care for acute myocardial infarction (AMI) and stroke. Their findings were such that, based on the estimated change in intercept and slope from pre to post intervention using segmented regression, they found insufficient evidence of a statistically significant effect on quality of care for stroke, although potential clinically important effects for AMI cannot be ruled out.

[8] used controlled interrupted time series analysis to quantify the effect of the introduction of 20 mph (32 km an hour) traffic speed zones on road collisions, injuries, and fatalities in London. The results were, the introduction of 20 mph zones was associated with a 41.9% (95% confidence interval 36.0% to 47.8%) reduction in road casualties, after adjustment for underlying time trends. The percentage reduction was greatest in younger children and greater for the category of killed or seriously injured casualties than for minor injuries. There was no evidence of casualty migration to areas adjacent to 20 mph zones, where casualties also fell slightly by an average of 8.0% (4.4% to 11.5%). Thus, they concluded that 20 mph zones are effective measures for reducing road injuries and deaths.

The significance of this inquiry is to determine if the policy of mandatory HIV/AIDS testing for all, has had any impact on the disease trend/pattern. This was done through development of forecasting models for predicting number of expected cases under this policy intervention so as to guide policy intervention and control measures on time. The rest of the paper is organized as follows: we discuss the methodology considered in Section 2. The results and discussion are contained in Section 3 and lastly, we conclude the paper in Section 4.

2. Methods and Materials

HIV/AIDS data for the study is taken from the administrative data submitted to the Copperbelt Provincial Medical Office (CPMO) as cases by 311 health facilities. The time series for monthly HIV/AIDS aggregated data is from 2010 to 2019 accounting for 120 observations and the annual summary of recorded cases is presented in Table 1. The sample size is 2842212. $X_{t=2010,\cdots,2019}$ where X_t represents monthly observations. MS EXCEL package was used as a primary data storage and EVIEWS version 9.0 software is used to implement stochastic models and graphical representations.

2.1. Stochastic Modelling

The Box-Jenkins methodology developed by [5] is a systematic process which is executed by using an iterative process until an adequate model is achieved. The

Table 1. Shows a summary of HIV/AIDS cases from the total samples tested annually.

Year	Tested	Positive
2010	130,977	37,365
2011	163,795	36,764
2012	184,209	35,251
2013	290,280	35,866
2014	376,395	37,003
2015	408,572	36,410
2016	480,166	33,155
2017	899,629	41,119
2018	867,026	44,466
2019	517,145	33,603

procedure is achieved by a step-by-step process of model; identification, specification, estimation, diagnostics and forecast. The ARIMA model has three parameters and these are; autoregressive (p), differencing order (d) and the order of moving average (q). The generic Box-Jenkins models are denoted by ARIMA (p, d, q) given by

$$\varnothing(B)(X_t - \mu) = \theta(B)e_t \tag{1}$$

where, $\varnothing, \theta$ and e_t are autoregressive parameter, moving average parameter and residual respectively. The residuals are assumed to be independent and identically distributed (i.i.d) normal random variables [9] [10].

2.2. Measure of Forecast Accuracy

The statistics are used to compare how well models fit the time series. Reference [11] explains that the best fit model is one with a high number of smaller errors. These errors are; the Root Mean Square Error (RMSE), the Mean Absolute Percentage Error (MAPE), the Mean Percentage Error (MPE) and the Mean Absolute Error (MAE). Forecast error is given by,

$$FE_t = O_t - F_t \tag{2}$$

where FE_t is the forecast error, O_t is the observed value and F_t is the forecast value.

The AIC and SBC are given by

$$\text{AIC} = 2\left(\log(L) - k\right) \tag{3}$$

$$\text{SBC} = 2\log(L) + k\log(n) \tag{4}$$

Other measures of forecast accuracy are given by

$$\text{MSE} = \frac{\sum_{t=1}^{n}(O_t - F_t)^2}{n} \tag{5}$$

$$\text{MAE} = \frac{\sum_{t=1}^{n}|FE_t|}{n} \tag{6}$$

$$\text{MAPE} = \frac{\sum_{t=1}^{n}\left|\frac{FE_t}{O_t}\right|}{n} \times 100 \qquad (7)$$

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, $\log(.) = \log_e(.)$ computes the natural logarithm and $t = 1, 2, 3, \cdots, n$.

3. Results and Discussion

This is the first step in the Box Jenkins procedure of modelling time series where the sample autocorrelation function (SACF) and sample partial autocorrelation function (SPACF) are used as guides to choosing the appropriate Autoregressive Integrated Moving Average (ARIMA) model. The time plot for HIV/AIDS monthly cases is displayed in **Figure 1**.

Figure 2 displays the two graphical tools which are the autocorrelation function (ACF) and the partial autocorrelation function (PACF), and their plots shows that the ACF decays exponentially and the PACF has a single spike at lag 1 indicating that the series is generated by an ARIMA (1, 0, 0) process.

Figure 3 shows that the Dickey-Fuller test statistic (4.270636) which is greater than (2.886074 at 5% level of significance). Hence, we reject the NULL hypothesis and conclude that the series has no Unit Root (the series is stationary).

3.1. Parameter Estimation

Figure 4 shows the estimated model parameters at 5% level of significance. The autoregressive (AR) parameter is statistically significant implying that the parameter influenced the model and no other. The mathematical representation of
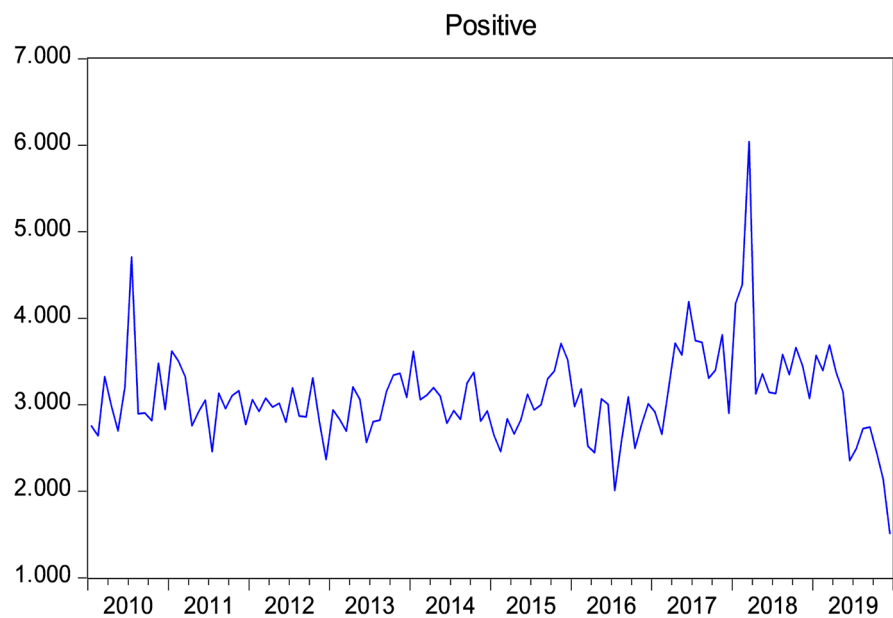

Figure 1. Time plot for monthly HIV/AIDS cases in Zambia from 2010 to 2019.

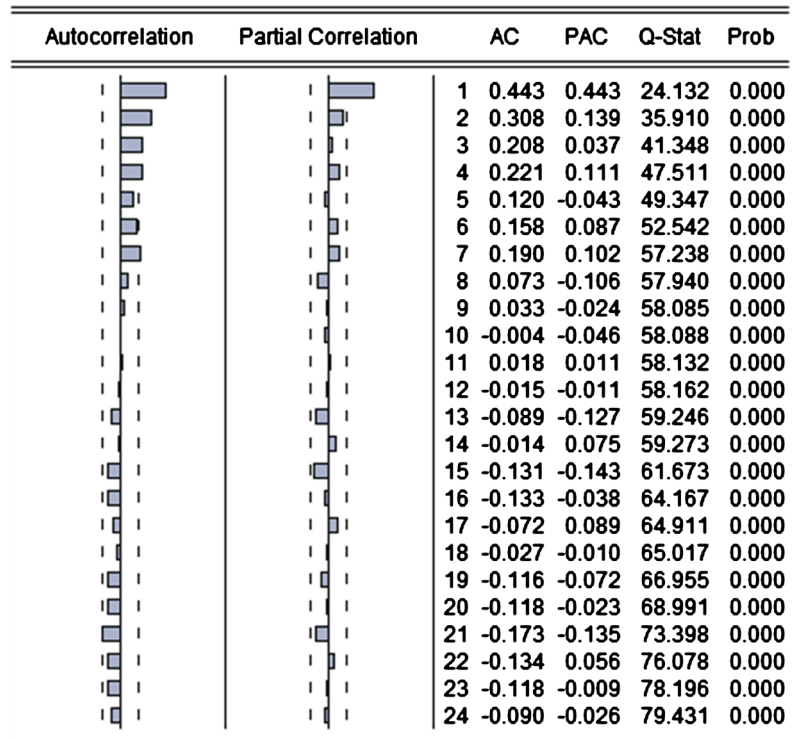| Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.443	0.443	24.132	0.000
		2	0.308	0.139	35.910	0.000
		3	0.208	0.037	41.348	0.000
		4	0.221	0.111	47.511	0.000
		5	0.120	-0.043	49.347	0.000
		6	0.158	0.087	52.542	0.000
		7	0.190	0.102	57.238	0.000
		8	0.073	-0.106	57.940	0.000
		9	0.033	-0.024	58.085	0.000
		10	-0.004	-0.046	58.088	0.000
		11	0.018	0.011	58.132	0.000
		12	-0.015	-0.011	58.162	0.000
		13	-0.089	-0.127	59.246	0.000
		14	-0.014	0.075	59.273	0.000
		15	-0.131	-0.143	61.673	0.000
		16	-0.133	-0.038	64.167	0.000
		17	-0.072	0.089	64.911	0.000
		18	-0.027	-0.010	65.017	0.000
		19	-0.116	-0.072	66.955	0.000
		20	-0.118	-0.023	68.991	0.000
		21	-0.173	-0.135	73.398	0.000
		22	-0.134	0.056	76.078	0.000
		23	-0.118	-0.009	78.196	0.000
		24	-0.090	-0.026	79.431	0.000

Figure 2. ACF and PACF for HIV/AIDS cases.

Null Hypothesis: Test Positive has a unit root
Exogenous: Constant
Lag Length: 1 (Automatic - based on AIC, maxlag=12)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-4.270636	0.0008
Test critical values:	1% level	-3.486551	
	5% level	-2.8860	
	10% level	-2.579931	

*MacKinnon (1996) one-sided p-values.

Figure 3. Augmented Dickey Fuller test (stationary test).

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	11.06764	33.10133	0.334356	0.7389
AR(1)	-0.361073	0.066811	-5.404378	0.0000
SIGMASQ	16882.9	16795.15	10.05546	0.0000
R-squared	0.133110	Mean dependent var		10.61538
Adjusted R-squared	0.113408	S.D. dependent var		443.8234
S.E. of regression	417.8998	Akaike info criterion		14.94231
Sum squared resid	15368340	Schwarz criterion		15.02508
Log likelihood	-676.8750	Hannan-Quinn criter.		14.97570
F-statistic	6.756151	Durbin-Watson stat		2.274073
Prob(F-statistic)	0.001864			
Inverted AR Roots	-36			

Figure 4. Estimated parameters and significant tests.

the AR (1) model is

$$(1-\varnothing B)(X_t - \mu) = e_t,$$

$$X_t = \mu + \varnothing(X_{t-1} - \mu) + e_t \tag{8}$$

From Equation (8), we estimate the parameter $\hat{\varnothing}$ by the method of moments using Yule Walker estimator for $\hat{\varnothing}$ which is $\hat{\varnothing} = \hat{\rho}_1$. Thus, $\hat{\varnothing} = -0.3816$.

The moment estimator for μ is $\bar{X}_t$ given by

$$\sum_{t=1}^{120} \frac{X_t}{n} = \frac{371003}{120} = 3091.68$$

Thus, the model is

$$X_t = 3091.68 - 0.3816(X_{t-1} - 3091.68) + e_t \tag{9}$$

3.2. Diagnostic Checks

Figure 5 shows the Box-Djung statistics is significant at all lags and the values of $\hat{\rho}_k$ or $\hat{\varnothing}_k$ are all lying within the range of $\frac{\pm 2}{\sqrt{n}}$, hence we have insufficient evidence to reject H_o (the process is a white noise) and conclude that the model assumptions are indeed that of a white noise.

In Figure 6, the histogram shows that the average of residuals is approximately 0. And that the (Jarque-Bera) normality test of residues is statistically significant at 5% level of significance. Thus, we conclude that the residues are normally distributed.

3.3. Forecasting

Box-Jenkins approach to forecasting stationary time series is reasonably stress-free.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.015	-0.015	0.0212	
		2	-0.017	-0.018	0.0500	0.823
		3	0.164	0.163	2.6613	0.264
		4	0.124	0.132	4.1835	0.242
		5	-0.245	-0.244	10.129	0.038
		6	0.060	0.032	10.496	0.062
		7	0.045	0.007	10.698	0.098
		8	-0.101	-0.044	11.749	0.109
		9	-0.114	-0.077	13.105	0.108
		10	-0.171	-0.275	16.178	0.063
		11	-0.011	0.033	16.190	0.094
		12	-0.120	-0.067	17.753	0.087
		13	-0.124	-0.094	19.438	0.079
		14	0.029	0.046	19.534	0.107
		15	-0.075	-0.181	20.170	0.125
		16	-0.039	0.062	20.344	0.159
		17	0.029	-0.005	20.444	0.201
		18	0.116	0.060	22.017	0.184
		19	0.028	0.082	22.108	0.227
		20	0.207	0.097	27.238	0.099
		21	0.003	-0.024	27.239	0.129
		22	-0.007	-0.094	27.246	0.163
		23	0.044	-0.023	27.489	0.193
		24	0.055	0.038	27.868	0.221

Figure 5. Serial correlation test (Box-Djung test) for the series.

The *k* step ahead forecast value X_{t+k} given all observations up to *n*, is denoted by $\hat{X}_t(k)$.

Figure 7 shows two trajectories of forecasted HIV/AIDS cases and actual cases of 18 months from 2017 to 2019 using ARIMA (1, 0, 0). Their behaviour is such that the predicted cases have a steady decline pattern compared to the actual cases which exhibits a sudden downward decline in the pattern after the intervention.

Figure 8 shows forecast error bounds, with a display of measures of forecast accuracy which were considered for AR (1) model goodness of fit. In the same setting the bias proportion tells us how far the mean of the prediction is from the mean of the actual series, the variance proportion tells us how far the variation of the forecast is from the variation of the actual one and finally the covariance proportion measures the remaining unsystematic forecasting errors. Note that the bias, variance, and covariance proportions add up to one.

3.4. Discussion

Our future ability to control HIV is dependent on the skills and proficiency of
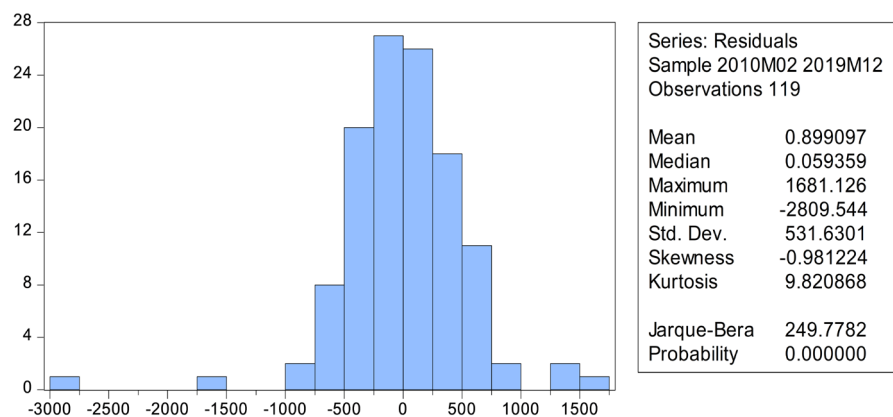

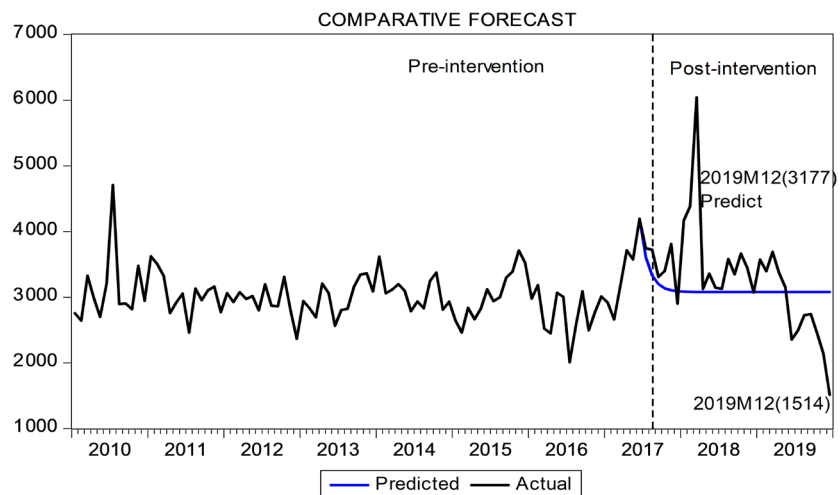**Figure 6.** Histogram and normality test of residuals

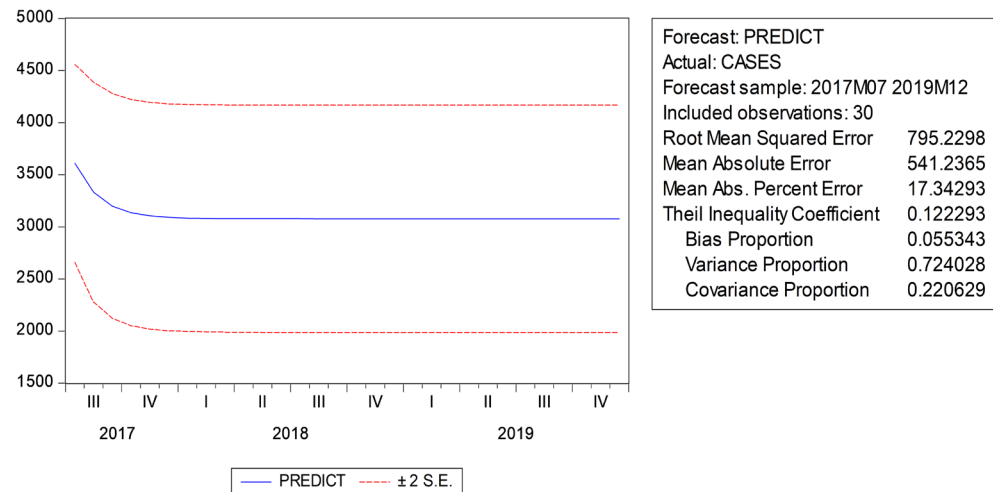**Figure 7.** Trajectories of forecasted HIV/AIDS cases and actual ones.

Figure 8. Forecast error bounds, with a display of measure of forecast accuracy, and a decomposition of the mean squared forecast error.

scientists, health care providers and public health professionals working in all parts of Zambia. HIV/AIDS is a huge public health problem with respect to morbidity and its burden on health care facilities. Persons living with HIV/AIDS are vectors of the disease and potential dangers to the community. In addition, persons living with the virus face a lot of stigma and discrimination from the general populace. Which result in people shunning away from getting tested for the said virus. This norm poses a challenge towards the fight against its spread in as much as the carriers themselves are not accounted for. Thus, planning and budgeting for HIV/AIDS related activities is a challenge. However, the government of the republic of Zambia saw it fit to make it mandatory to test for HIV/AIDS virus whenever other illnesses are attended to.

The ARIMA (1, 0, 0) is a suitable model which best fits the HIV/AIDS time series cases and is therefore appropriate for forecasting HIV/AIDS cases per month in the future based on observed HIV/AIDS cases in past times. The results also review that in the last 10 years the HIV/AIDS pattern has been a fluctuating trend, with a highest peak being recorded in the late 2017 to early 2018 just after policy implementation. Then cases declined below average of around 14.29% as forecasted to a record breaking nose dive of 57.4%, towards the end of the year 2019. The potential implication of this study is that by developing forecasting models for predicting HIV/AIDS cases in advance on a regular basis is to support internal decisions and planning for programs such as test and treat.

4. Conclusion

In this paper, the Box-Jenkins modelling process is used to determine an ARIMA model and advance to forecasting. The HIV/AIDS data for the study were taken from Ministry of Health (Copperbelt province administrative office) for the period, 2010 to 2019 all-ages inclusive, with 120 observations. The results show that ARIMA (1, 0, 0) is a suitable model which best fits the HIV/AIDS time se-

ries cases and is therefore appropriate for forecasting. The model predicts a reduction from 3500 to 3177 representing 14.29% in HIV/AIDS cases from 2017 (year of policy activation) to 2019, but the actual recorded cases dropped from 3500 to 1514 accounting for 57.4% in the same time frame. Finally, the robust efforts being made by the Zambian government aimed at achieving the 90-90-90 target of ending AIDS by 2030 are indeed a workable solution. But more needs to be done such as behavioural change sensitizations and also implementing programmes aimed at reducing risk behaviours.

Acknowledgements

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1]   World Health Organization (2020) World HIV/AIDS Report. Geneva, Switzerland. https://www.who.int/news-room/fact-sheets/detail/hiv-aids

[2]   Zambia Statistical Agency (2016) Zambia Population HIV Impact Assessment (Zamphia) Survey. https://www.zamstats.gov.zm/

[3]   Unaids and Joint United Nations Programme on HIV/AIDS (2010) Getting to Zero: 2011-2015 Strategy. World Health Organization.

[4]   Brodersen, K.H., Gallusser, F., Koehler, J., Remy, N. and Scott, S.L. (2015) Inferring Causal Impact Using Bayesian Structural Time-Series Models. *The Annals of Applied Statistics*, **9**, 247-274. https://doi.org/10.1214/14-AOAS788

[5]   Box, G.E., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015) Time Series Analysis: Forecasting and Control. John Wiley & Sons, Hoboken.

[6]   Ma, Z.Q., Kuller, L.H., Fisher, M.A. and Ostroff, S.M. (2013) Peer Reviewed: Use of Interrupted Time-Series Method to Evaluate the Impact of Cigarette Excise Tax Increases in Pennsylvania, 2000-2009. *Preventing Chronic Disease*, **10**. https://doi.org/10.5888/pcd10.120268

[7]   Taljaard, M., McKenzie, J.E., Ramsay, C.R. and Grimshaw, J.M. (2014) The Use of Segmented Regression in Analysing Interrupted Time Series Studies: An Example in Pre-hospital Ambulance Care. *Implementation Science*, **9**, Article No. 77. https://doi.org/10.1186/1748-5908-9-77

[8]   Grundy, C., Steinbach, R., Edwards, P., Green, J., Armstrong, B. and Wilkinson, P. (2009) Effect of 20 mph Traffic Speed Zones on Road Injuries in London, 1986-2006: Controlled Interrupted Time Series Analysis. *BMJ*, **339**, b4469. https://doi.org/10.1136/bmj.b4469

[9]   William, W.S. and Wei, S. (2006) Time Series Analysis: Univariate and Multivariate Methods. USA, Pearson Addison Wesley, Segunda Edicion, 212-235.

[10]   Jere, S. and Moyo, E. (2016) Modelling Epidemiological Data Using Box-Jenkins

Procedure. *Open Journal of Statistics*, **6**, 295-302.
https://doi.org/10.4236/ojs.2016.62025

[11] Siluyele, I. and Jere, S. (2016) Using Box-Jenkins Models to Forecast Mobile Cellular Subscription. *Open Journal of Statistics*, **6**, 303-309.
https://doi.org/10.4236/ojs.2016.62026