Time Series Forecasting of Diagnosed HIV Cases in Davao Region using ARIMA Models

Jamaiya Hadji Acmad | Jecelle Israel | Judilee John Ranara

PROBLEM

Introduction

This study explores the application of AutoRegressive Integrated Moving Average (ARIMA) modeling technique for time series forecasting of HIV cases in Davao Region, Philippines. The findings of the study can contribute to informed decision-making and resource allocation for effective healthcare planning and intervention strategies towards HIV prevention and control in the region.

Background of the Study

The rising prevalence of Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome (HIV/AIDS) poses a significant public health threat globally, including in the Philippines. Compared to the other regions in the Philippines, Region XI ranked 5th, contributing to 1,347 new cases in the entire country from January to September 2022 (Cayon, 2022), with monthly cases not less than 50. With the aforementioned problem, there is a need and urgency to address it and develop effective strategies for prevention and control. Time series analysis and forecasting are crucial for predicting future trends, with the use of ARIMA models. The ARIMA modeling technique has been widely employed in numerous studies to forecast HIV cases accurately. Thus, this study is designed to develop a reliable forecasting model for predicting monthly diagnosed HIV cases in Region XI for a year. By doing so, proactive planning and response can be facilitated, enabling healthcare systems to meet the growing demand for HIV-related services effectively in terms of timely implementation of control measures and interventions.

DATA

Data Preprocessing

The dataset was initially collected through business correspondence with The Regional HIV/STI Surveillance Unit of the Davao Center for Health Development. One of the researchers sent an email approved by the course instructor to the said office requesting for historical data of number of monthly diagnosed HIV cases in Davao Region for the year 2003-2022. The data transformation revolves on stabilizing variance as applying a logarithmic transformation, the variability can be stabilized, allowing for more reliable statistical analysis and modeling.

TECHNIQUES

Feature Construction

As shown in Figure 1, the monthly newly diagnosed cases of HIV in the Philippines from January 2003 to December 2022 was made into a time series plot with a frequency of 12. Since stationarity is an important assumption for ARIMA models, there is a need to determine if the data is stationary. The initial inspection of the original data conducted through plotting time series decomposition, shows increasing trend and no seasonality where data exhibits long-term increase and no seasonal behavior. Hence, the resulting time series plot of the data is said to be nonstationary. However, employing a unit root test was viewed as a more effective approach to assess the stationarity of the data.

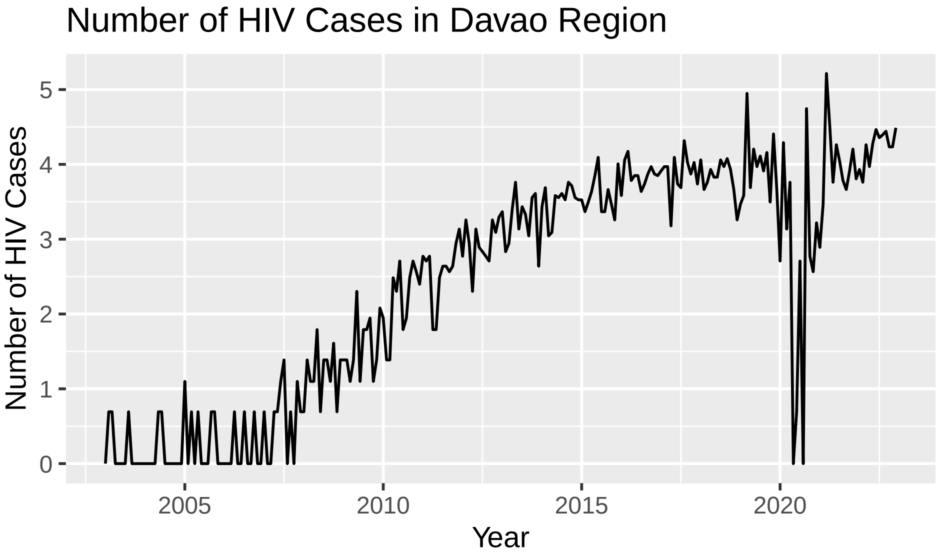

Figure 1: Time Series Plot of HIV Cases in Davao Region

Feature Selection and Transformation

The researchers then used the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. The test statistic **(4.284)** is much bigger than the 1% critical value **(0.739)**. Hence, the null hypothesis was rejected which concludes that the original data are not stationary. This implied the need for differencing the data and apply the test again.

After differencing the original data and subsequently subjecting it again to the KPSS test, the data became stationary, with the test statistics **(0.0253)** which is smaller than the critical value **(0.739)**. The function **ndiffs()** was utilized to support the claim of how many times our data must be differenced, indicating **d=1**. In ARIMA modeling technique, the “d” component refers to the differencing parameter, which represents the number of times differencing is applied to the time series data to make it stationary.

Model Comparison

First, the **auto.arima()** function with no seasonality was used to automatically fit an ARIMA model. Then, the resulting ARIMA model was **MA(0,1,1)** with drift.

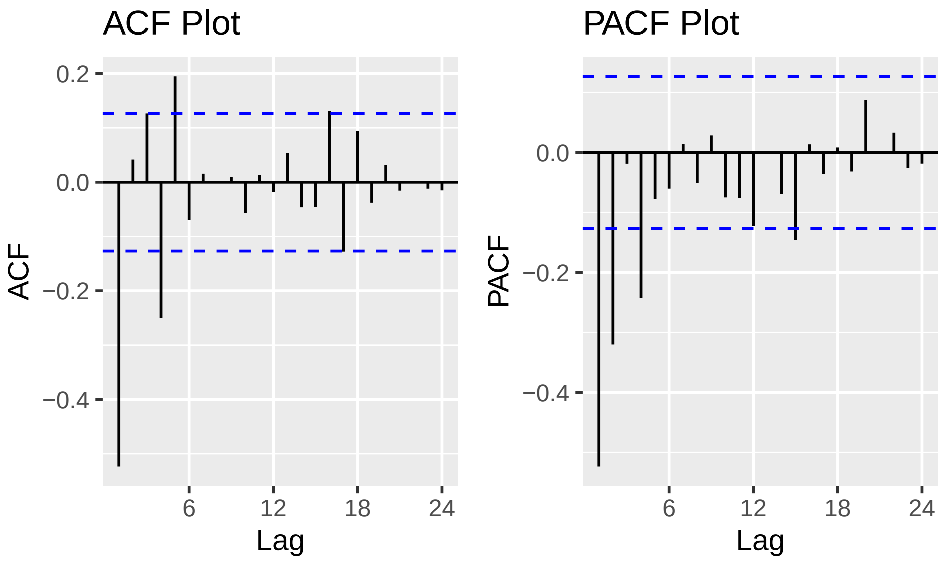

Figure 2: ACF and PACF Plots of the Differenced Data

Next, the researchers let the PACF and ACF plots of the differenced data to determine the values of p and q, respectively. As shown above, the PACF shows exponentially decaying or sinusoidal pattern. While the ACF plot contains two significant spike, the PACF plot has three significant spikes and then almost no spikes thereafter. Consequently, the ACF and PACF lead the researchers come up with **ARIMA(0,1,2)** model and deemed as appropriate one.

Then, the researchers improve the model by incorporating arguments **stepwise = FALSE** and **approximation = FALSE** to **seasonal = FALSE** in automatically fitting another ARIMA model. After which, the model generated was **ARIMA(3,1,2)** with drift. Consequently, the researchers decided to consider again another model, but this time, it was an autoregressive model which is **AR(3,1,0)**, as the higher the **p**, the data will be more smoothed.

Table 1: The AICc values of the four (4) ARIMA models

NO.	Model	AICc
1	ARIMA(0,1,2)	394.29
2	ARIMA(3,1,0)	400.81
3	ARIMA(0,1,1) w/o Seasonality	392.56
4	ARIMA(3,1,2) w/o Seasonality, approximation, stepwise	387.76

The Akaike Information Criterion with a correction (AICc) was utilized as measure to compare different ARIMA models with the different combinations of parameters. From the Table 1, the **auto.arima()** function w/o seasonality, approximation, stepwise generated a model that achieved the lowest AICc among the four models evaluated which is **(387.76)**. In this case, the model

ARIMA(3,1,2) was considered for the forecasting process, as it demonstrates strong performance based on the AICc criterion.

Model Training and Validation

Root Mean Squared Error (RMSE) is a metric used to evaluate the accuracy of a predictive model by measuring the average magnitude of the prediction errors. A lower RMSE indicates smaller prediction errors and generally suggests a more accurate model in terms of point predictions.

Table 2: The RMSE values of the four (4) ARIMA models

NO.	Model	RMSE
1	ARIMA(0,1,2)	0.5431849
2	ARIMA(3,1,0)	0.5484074
3	ARIMA(0,1,1) w/o Seasonality	0.5411318
4	ARIMA(3,1,2) w/o Seasonality, approximation, stepwise	0.5261236

Among the four models, it was observed that the model **ARIMA(3,1,2)** exhibited the smallest RMSE which is **(0.5261236)**. Thus, the researchers considered this model for data forecasting.

Before selecting a model to forecast, all models were subjected to model training and fitting. This process is crucial as it helps us see how the proposed model is adaptive or sensitive to changing data patterns, which allows us to optimize more the model parameters and evaluate the model performance. Based on the time series, there can be seen a sudden downward shift of data from some months in the year 2020. Upon checking the data, there were months in the said year where no cases are recorded, which can be explained by some factors. Consequently, the models were fitted from the starting year until 2019.

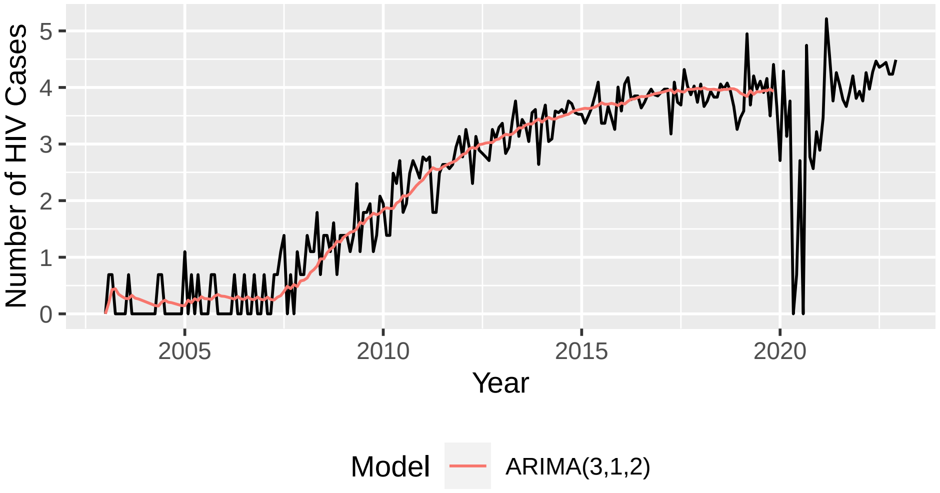

Figure 3: Model Fitting Plot of the Selected Model ARIMA(3,1,2)

Among the four models, the best fitted model is **ARIMA(3,1,2)** with drift, which is the model we considered earlier as it has indicators of being a good model (with lowest AICc and smallest RMSE).

RESULTS AND DISCUSSION

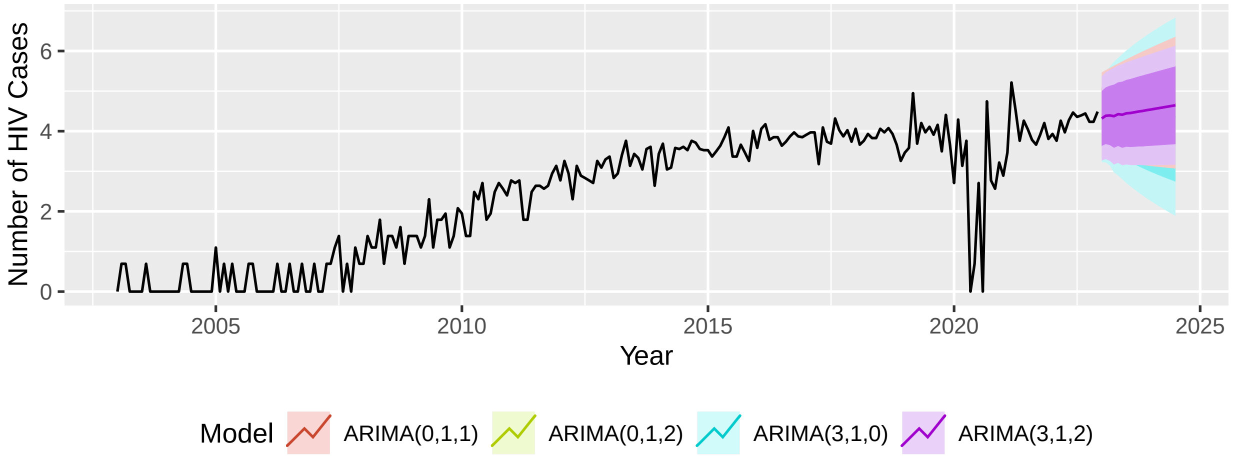

Figure 4: Time Series Plot of Forecasted HIV Cases in Davao Region from January 2023 to July 2024

All models were also applied to forecast the HIV cases in the region for the next 19 months, spanning from January 2023 to July 2024. Among the four models, the lowest AICc and RMSE values certainly point on a specific model which is **ARIMA(3,1,2)**, which is also the best fit model. With this, the researchers suggested that this model will be the most accurate compared to the other models.

Based on the time series plot with the forecasted data from and focusing on the selected model, the forecast shows that the HIV cases in Davao Region will continue to increase, exhibiting an increasing trend.

CONCLUSION

In light of the findings, the time series forecasting of HIV cases in Davao Region using ARIMA models indicates an increasing trend. The forecasted data consistently demonstrates a rising pattern in the number of HIV cases over the study period. The results of this endeavor underscores an important **insight** which is the urgent need of proactive measures and targeted interventions to address the increasing burden of HIV in the region, particularly widespread sex education, and accessible healthcare services. Furthermore, the researchers conducted this study of time series analysis forecasting of HIV cases in Davao City using ARIMA models for a cause which presents benefits to various domains of human existence.

First, **policymakers and healthcare authorities** can use these accurate forecasts to allocate resources, plan targeted interventions, and implement control measures to combat HIV spread.

Healthcare providers can benefit by anticipating HIV-related services demand, ensuring timely and efficient care.

Additionally, **HIV/AIDS organizations and NGOs** can optimize their outreach and intervention efforts, targeting high-risk areas and populations.

Researchers and the academe can also leverage the forecasting model to advance studies in public health and epidemiology.

Lastly, the **general public** can benefit from increased awareness, informed sexual health decisions and preventive measures.

RECOMMENDATIONS

The researchers found out that the more models to be considered (meaning: the more differently-arranged combination of values for (p,d,q) parameters), the higher the possibility for them to acquire a model that is best fit, with lowest AICc, or smallest RMSE. However, due to time constraints, the researchers only consider few combinations, which are based on **auto.arima()** functions and ACF and PACF plots. Moreover, the historical data can also be detrended and smoothed, instead of differencing, considering that the data has no seasonality patterns.

Based on the results of the study, and the conclusions drawn from them, the following recommendations were suggested. The proponents of this project recommend the future researchers to address the limitations of the study and to construct the same research in a new context or location. Future studies can also consider utilizing other forecasting techniques that can be suited for the type of the data.

CODES

The time series analysis and forecasting of the historical data were performed using RStudio. The codes and raw dataset used can be accessed through this link: https://github.com/upminruru/AMAT132-Final-Project-FILES-.git.

REFERENCES

Apa-ap R. & Tolosa, H. (2018). Forecasting the Monthly Cases of Human Immunodeficiency Virus (HIV) of the Philippines. *Indian Journal of Science and Technology.* Polytechnic University of the Philippines, College of Science, Department of Mathematics and Statistics. 11(47), doi: 10.17485/ijst/2018/v11i47/121923.

Cayon, C. (2022). HIV cases increasing in Davao Region. *Philippine Information Agency.* Retrieved on June 12, 2023 from https://pia.gov.ph/news/2022/12/19/hiv-cases-increasing-in-davao-region.

Kurniasari, M., Huruta, A., Tsai, H., Lee, C. (2021). Forecasting future HIV infection cases: evidence from Indonesia. *Soc Work Public Health.*36(1):12-25. doi: 10.1080/19371918.2020.1851332.

Yang, Y., Zhu, Y., Tseng, S., Tang, L., Chen, Y. & Guo, X. (2021). Prediction and analysis of HIV/AIDS incidence based on ARIMA model in China. *29th International Conference on Orange Technology (ICOT).* Tainan, Taiwan. pp. 1-4, doi: 10.1109/ICOT54518.2021.9680664.