Time Series Modeling of Philippine Political Violence Fatalities

Daniel Dave Cruz, Pamela Ann Gloria, Andrea Nicole Senson

Abstract

Despite being a democratic country, the Philippines is still riddled with issues of political violence. These involve controversial issues such as EJKs, massacres, and insurgent conflicts, harming the lives of many Filipinos. This paper aims to produce a time series model for political violence fatalities in the Philippines in order to help understand the underlying trends and make policy assessment regarding conflict events more efficient.

The scope of the dataset from The Humanitarian Data Exchange is from January 2016 to May 2024. Preprocessing steps involved log transformation, followed by backward first-order differencing to assure stationarity and autocorrelation. Eleven candidate models were generated by analyzing the cut-offs and spikes of the correlograms of PACF and ACF. Using Akaike Information Criterion (AIC) scores to choose the model, we end up with the ARMA(6, 1) process $Y_t = -0.0320 + 0.1985Y_{t-1} - 0.0027Y_{t-2} + 0.0338Y_{t-3} - 0.1066Y_{t-4} - 0.1463Y_{t-5} - 0.3384Y_{t-6} + Z_t - 0.6863Z_{t-1}$ as our best model. An analysis of the coefficients in the model suggests that a gradual increase from a small number of fatalities or consecutive months with dips would produce high predicted spike Y_t in fatalities.

1 Introduction

1.1 Background of the Study

Philippine politics is influenced by an intricate interplay of cultural, historical, and social factors, which is rooted in the country's complex history of revolution and colonization [7]. It is concerned about the acquisition and use of power as authorities from the local barangay level up to the highest levels of government compete for political positions. The political structure is primarily democratic, and in spite of the country's attempts for economic reform, the Philippine government is still overpowered by issues of poverty, corruption, and crime [8].

Given the country's complex political background, political violence emerges as an important and concerning reality. By definition, political violence is the term for hostile and aggressive acts motivated by political goals, such as a desire to influence political change or bring about modifications to governance [2]. This includes a variety of activities, such as the infamous extrajudicial killings (EJKs), violence during elections, and the ongoing conflict between government and insurgent bodies. For instance, during the reign of President Rodrigo Duterte, 29,000 people were killed by EJK [6]. The Maguindanao massacre last 2009 stands as the world's deadliest single-day assassination of 58 media workers [3]. Lastly, the Moro conflict in Mindanao, which includes the Moro Islamic Liberation Front and the Moro National Liberation Front , has resulted in fatalities and displacement of communities [5]. Clearly, political violence is a pressing concern in the Philippines, impacting many Filipinos to their detriment.

Submission for MATH 62.2 Time Series and Forecasting, AY 2023-2024 Second Semester

1.2 Significance of the Study

This paper seeks to analyze the trends of political violence fatalities in the Philippines by producing a linear time series model that fits the collected data sufficiently. The said model can then give insightful data on anticipated political violence scenarios in the country, which would be crucial for the government to make efficient policy-making decisions. Authorities can conduct a comprehensive evaluation of the effectiveness of existing conflict interventions and modify them based on the changes in the trends of fatalities.

Hence, the linear time series model gives a methodical framework for understanding and handling political violence, which could then in attaining justice, peace, and stability in the Philippines.

1.3 Scope and Limitations

This study will only aim to create a linear time series model for political violence fatalities in the Philippines. It will not perform forecasting and estimation, and it will not perform hypothesis testing to determine the significance of the coefficients of the resulting model. An in-depth analysis on the factors affecting the numbers will also not be performed. This means that the interpretation of the model will only revolve around the effect of each component to the model in its entirety. Additionally, the methods used in this study do not guarantee that we will arrive at the best possible model, but will give a model that is a relatively good fit for the collected data. Furthermore, the goal of the paper is to answer the problem: What are the trends in the occurrence of political violence in the Philippines?

2 Methods

2.1 Theoretical Framework

A time series $\{X_t\}$ is defined as a set of observations of the same random variable recorded over different points of time [4]. One such practical application of interest to people is the modeling of time series data. This helps understand patterns and trends in the data, and make predictions of potential future values.

One such class of models is the linear process, which has the representation

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j},$$

for all t, where $Z_t \sim \text{WN}(0, \sigma^2)$ (i.e., Z_t is an uncorrelated time series each with mean 0 and variance σ^2) and $|\psi_j|$ is a sequence of constants such that $\sum_{j=-\infty}^{\infty} |\psi_j|$ is convergent. Under this class includes three truncated versions of the model [9].

First is the autoregressive process of order p, or AR(p), denoted by

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t,$$

where $Z_t \sim \text{WN}(0, \sigma^2)$ and Z_t is uncorrelated with $X_1, X_2, \ldots, X_t$. Second is the moving average process of order q, or MA(q), represented by

$$X_t = \mu + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$). The last is the autoregressive moving average (ARMA) with orders p and q, or ARMA(p, q), with representation

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \alpha + Z_t + \theta_1 Z_{t-1+\cdots+\theta_q Z_{t-q}},$$

where $Z_t \sim \text{WN}(0, \sigma^2)$ and $\alpha = \mu\,(1 - \phi_1 - \cdots - \phi_p)$. Our aim is to model our data using one of these three models. And in order to choose which is the best model, we use the Akaike Information Criterion (AIC), which is computed by AIC $= \frac{-2 \ln \text{likelihood}}{T} + \frac{2p}{T}$, where p is the number of model parameters, L is the likelihood function, and T is the sample size [10]. The lower the AIC, the better the model fit.

But before modeling, the time series data must satisfy two conditions. First, it must be stationary. Specifically, we want the data to be weakly stationary, implying that the mean of any data point

$\mathbb{E}(X_t)$ is independent of t, and that the autocovariance at lag h $\gamma_X(t+h,t)$ is independent of t for any value of h. The purpose of doing so is to ensure that something within the data does not vary with time, hence is something we can predict.

Stationarity can be checked using the Augmented Dickey-Fuller (ADF) Test. This test assesses whether the time series possesses a unit root, indicative of non-stationarity by signifying the convergence of the linear process. Hence, if the ADF test is rejected, then the dataset is stationary, and vice versa. In the event that it fails, we can use differencing to create a new time series $\{\nabla X_t\}$, where $\nabla X_t = X_t - X_{t-1}$.

The second condition is that the data must possess autocorrelation. Autocorrelation (ACF) at lag h, or $\rho_h = \text{Cor}(X_{t+h}, X_t)$, is used to identify whether there is some relationship between time series data points at different sets of time. Without this, modeling is senseless.

This is checked using the Ljung-Box Test, which checks whether the data has no serial autocorrelation. If this is rejected, then serial correlation is present, hence modeling can be done.

2.2 Data Collection

The CSV dataset retrieved from The Humanitarian Data Exchange [1] consists of the total number of reported political violence events and fatalities in the Philippines from January 2016 to May 2024. The data is organized on a monthly basis, bringing the total number of instances to 101.

2.3 Assumptions

Suppose that the 'Fatalities' column of the dataset is $\{X_t\}$ with time step t in terms of months. For this study, we will assume that there are no features or any other characteristic in $\{X_t\}$ that would complicate its modeling, aside from trend and seasonality. However, we will assume the if seasonality is present, it can be dealt with via backwards first-order differencing. We will also assume that the trend is linear, hence justifying the use of truncated linear processes. Finally, this study will be using $\alpha = 0.05$ as the level of significance.

2.4 Data Exploration

All data exploration and modeling were done using R, with the libraries TSA and tseries. The code can be seen in the Appendices. In order for modelling to proceed, we must check for both stationarity and autocorrelation. A plot of the data is shown in Figure 1.

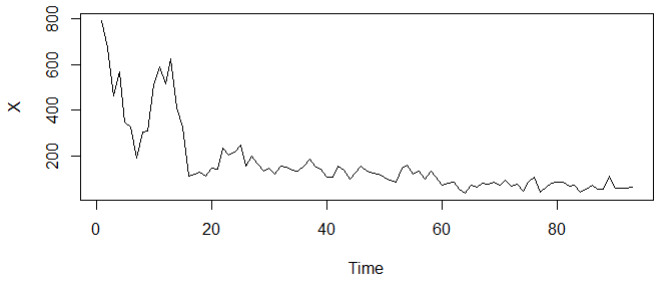

Figure 1: Plot of $\{X_t\}$.

Upon observation, the values in the plot are quite high, indicating the need for log transformation. Let us denote the log transformation of our data as X'_t.

Now, we test for stationarity using the ADF Test with H_0 : There is a unit root for $\{X'_t\}$ .vs. H_1 : There is no unit root for $\{X'_t\}$. We get a p-value less than $0.01 < \alpha$, hence we reject H_0. Thus, $\{X'_t\}$ is stationary.

Next, we see if the data has serial correlation via the Ljung-Box Test with H_0 : There is no autocorrelation .vs. H_1 : There is autocorrelation $\{X_t'\}$. Doing so, we obtain a p-value of $2.2 \times 10^{-16} < \alpha$, making us reject H_0. Thus, $\{X_t'\}$ has autocorrelation, hence we can proceed with modeling.

2.5 Modeling

We first generate a correlogram of the ACF to identify candidate MA models. However, the correlogram, as seen in Figure 2, suggests the presence of seasonality. Thus, we use backwards first-order differencing on the data. Let our new data be $\{Y_t\}$, such that $Y_t = X_t - X_{t-1}$.

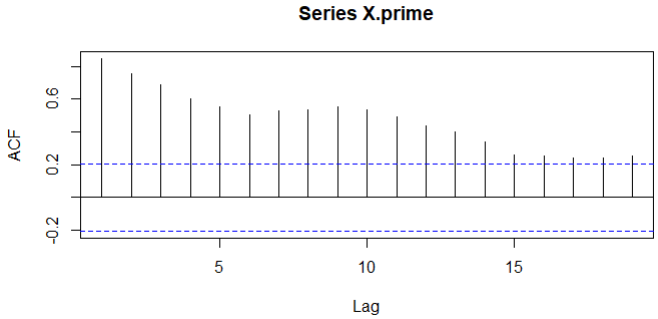

Figure 2: Correlogram for ACF of $\{X_t'\}$.

Once again generating a correlogram of ACF, we see, in Figure 3, that there is a cut-off at lag-1, but another sudden significant value at lag-6. We decided to take both MA(1) and MA(6) as candidates.

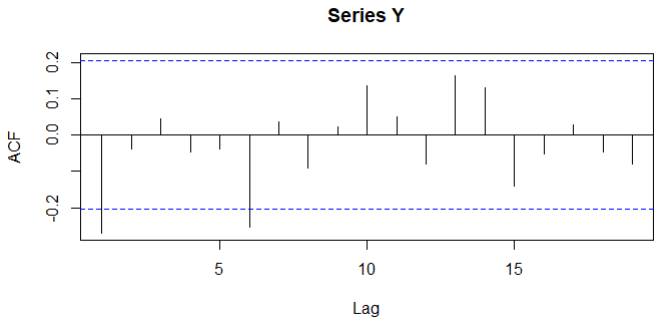

Figure 3: Correlogram for ACF of $\{Y_t\}$.

Afterwards, we create the correlogram for PACF (shown on Figure 4), in which we see another cut-off at lag 1, but sudden spikes at lags 6 and 8. Hence, our decision is to take all of AR(1), AR(6), and AR(8) as candidates for the AR model. Overall, we will also test ARMA(1, 1), ARMA(1, 6), ARMA(6, 1), ARMA(6, 6), ARMA(8, 1), and ARMA(8, 6).

3 Results and Discussion

3.1 Model Selection

In order to determine which of the 11 candidate models is the best fit for the data, we look at their AIC scores. The AIC of each model is summarized in Table 1.

It can be observed that the model with the lowest AIC, and thus the best model, is ARMA(6, 1). To write the model in functional form, we first solve for α.

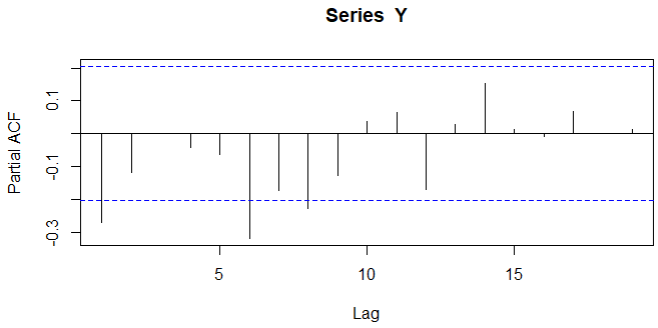

Figure 4: Correlogram for PACF of $\{Y_t\}$.

Table 1: Summary of AIC scores

Model	AIC Score
AR(1)	46.27
AR(6)	42.24
AR(8)	36.39
MA(1)	44.95
MA(6)	38.02
ARMA(1, 1)	39.36
ARMA(6, 1)	35.16
ARMA(8, 1)	36.79
ARMA(1, 6)	38.47
ARMA(6, 6)	36.92
ARMA(8, 6)	39.67

From Table 2, we know that the intercept $\mu = -0.0235$. Hence, we can solve for

$$\begin{aligned}\alpha &= \mu(1 - \phi_1 - \phi_2 - \phi_3 - \phi_4 - \phi_5 - \phi_6) \\ &= -0.0235(1 - 0.1985 - 0.0027 + 0.0338 - 0.1066 - 0.1463 - 0.3384) \\ &= -0.0320\end{aligned}$$

Therefore, the model for ARMA(6, 1) is represented in functional form as

$$\begin{aligned}Y_t = -0.0320 + 0.1985Y_{t-1} - 0.0027Y_{t-2} + 0.0338Y_{t-3} - 0.1066Y_{t-4} \\ -0.1463Y_{t-5} - 0.3384Y_{t-6} - 0.6863Z_{t-1} + Z_t.\end{aligned}$$

3.2 Model Interpretation

Note that since we're dealing with differenced data, it's better to interpret the model in terms of spikes and dips; that is, a negative value for Y_t would indicate a dip and a positive value for Y_t would be a spike in the number of fatalities from time step $t-1$ to t.

Interpreting the model component-wise, we know that a positive coefficient in the AR component means that an increase in its corresponding variable—a past difference—leads to a higher predicted change in fatalities at time t. On the other hand, a negative coefficient in the AR component suggests an inverse relationship. As its corresponding variable decreases, Y_t increases. For the MA component/s, a negative coefficient indicates that an increase in past white noise differences reduces Y_t.

It is also worth noting that the smallest coefficient in the entire model is 0.0027, which is for Y_{t-2}. While it may suggest low significance in determining Y_t, we cannot be certain since we did not perform hypothesis testing. Likewise, the largest coefficient, 0.3884 for lag-6, may be large but may also still be essentially zero.

Table 2: ARMA(6,1) Coefficients and Statistics

Coefficient	Estimate	Standard Error
ar1	0.1985	0.1359
ar2	-0.0027	0.1041
ar3	0.0338	0.0996
ar4	-0.1066	0.0996
ar5	-0.1463	0.1051
ar6	-0.3384	0.1110
ma1	-0.6863	0.1171
intercept	-0.0235	0.0067

What may be inferred from the model though is that the trajectory that would lead to high Y_t, which is the more alarming case, would be one where there is a gradual increase from dips (negative values) or very small spikes. This can be inferred from how the coefficients for Y_{t-6}, Y_{t-5}, and Y_{t-4} are negative and decreasing in magnitude, and from how the coefficients for Y_{t-3}, Y_{t-2}, and Y_{t-1} are generally positive. This means that if we see an increasing trend in the change in fatalities from 6 months ago to a month ago, we can expect a bigger spike or smaller dip in the current month.

Another notable feature of the model is its large, negative coefficient for Z_{t-1}. This implies that recent noise tends to dampen extreme changes in fatalities. If the previous error was large (positive or negative), the model adjusts the current value accordingly.

Moreover, it can be said that the intercept -0.0320 is the baseline change in fatalities when no other factors are considered. However, in real-world scenarios, the variables are rarely 0. Hence, the intercept should be considered alongside the other coefficients and variables.

3.3 Implications

The model suggests some implications for the country's policy-making bodies. While the system tends to self-correct, policymakers should still monitor and act when necessary. Since the trajectory for a higher Y_t is a steady increase in spikes, there must be an increased effort to monitor the spikes, and if consecutive increases persist, targeted interventions are needed.

Authorities must also set risk thresholds based on predicted changes from the model. If the risk exceeds a certain level, specific interventions, such as surveillance, can be triggered. This ensures efficient resource allocation since solutions to the country's political violence problem may be targeted towards those who are more at risk of having political violence events blow up.

References

[1] Philippines - Conflict Events.

[2] Violence.

[3] Timeline: The Maguindanao killings and the struggle for justice, Dec. 2019.

[4] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer Cham, 3 edition, Aug. 2016.

[5] E. Gutierrez and S. J. Borras. The Moro Conflict: Landlessness and Misdirected State Policies, Jan. 2004. ISBN: 9781932728149.

[6] H. Johnson and C. Giles. Philippines drug war: Do we know how many have died? *BBC*, Nov. 2019.

[7] C. J. Montiel. Philippine Political Culture and Governance. In *Philippine Political Culture: View from Inside the Halls of Power*, page 95. Philippine Governance Forum, 2002.

[8] S. Rogers. Philippine Politics and the Rule of Law. *Journal of Democracy*, 15(4):111–125, Oct. 2004.

[9] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications. With R Examples*. Springer Texts in Statistics. Springer Cham, 4 edition, Apr. 2017.

[10] R. S. Tsay. *An introduction to analysis of financial data with R*. John Wiley & Sons, 2014.

A Appendix

A.1 Raw Data

The dataset can be found at `https://data.humdata.org/dataset/philippines-acled-conflict-data?fbclid=IwAR32tczD2m_2evv_4liVa5OMMt77c0KNI7hdIui-I2x3InsjteCOG66PLcU`

A.2 R Code

```
# PRELIMINARIES
# Importing of Libraries
library(TSA)
library(tseries)

# Loading the data
# violence= read.csv("violence truncated.csv")
head(violence)
# We choose only the fatalities column
X = violence$Fatalities
X = ts(X)

# DATA EXPLORATION
# Transformation
plot(X) # Values extremely high. Need to log transform
X.prime = log(X)
plot(X.prime)
# Plot looks good now

# Autocorrelation
Box.test(X.prime, type="Ljung") # Has serial correlation
adf.test(X.prime) # Is stationary

# MODELING
# Acf
acf(X.prime) # Indicates possibility of seasonality
Y = diff(X.prime) # Differencing to address seasonality
```

```
plot(Y)
boxplot.stats(Y)
acf(Y) ##MA(1) or MA(6)

# Pacf
pacf(Y) #AR(1), AR(6) or AR(8)

# Creating the Models
# AR
arima(Y, order=c(1,0,0)) #AR(1) aic= 46.27
arima(Y, order=c(6,0,0)) #AR(6) aic= 42.24
arima(Y, order=c(8,0,0)) #AR(8) aic= 36.39
# MA
arima(Y, order=c(0,0,1)) #MA(1) aic=44.95
arima(Y, order=c(0,0,6)) #MA(6) aic=38.02
# ARMA
arima(Y, order=c(1,0,1)) #aic=39.36
arima(Y, order=c(1,0,6)) #aic=38.47
arima(Y, order=c(6,0,1)) #aic=35.16
arima(Y, order=c(6,0,6)) #aic=36.92
arima(Y, order=c(8,0,1)) #aic=36.79
arima(Y, order=c(8,0,6)) #aic=39.67

# FINAL MODEL IS ARMA(6,1)
```