ISSN 2250-3153

Out-of-sample forecasting of the Region XII, Philippines' non-metallics production volume using different modeling techniques

Sheila Mae D. Golingay

Department of Environment and Natural Resources, Mines and Geosciences Bureau XII,
City of Koronadal, South Cotabato, Philippines

DOI: 10.29322/IJSRP.10.08.2020.p10469
http://dx.doi.org/10.29322/IJSRP.10.08.2020.p10469

Abstract- This research paper investigates the accuracy of seven time series methods for short-term production volume forecasting. Different methods are compared to measure the forecasting performance on the monthly production volume of the Region XII, Philippines during the 2017-2019 period. The findings revealed that even the autoregressive integrated moving average (ARIMA) model performed well on the given data, but, overall, the best results were achieved with seasonal naïve method. This would lead us to conclude that even with little domain knowledge, simpler methods can outperform complex alternatives.

Index Terms- Philippines; Production; Forecasting; Mining; Natural Resources; Non-metallics

I. Introduction

One of the main indicators to measure the economic growth of a region in the Philippines is mineral production. The Board of the Philippine Statistics Authority (PSA) highlights the indicators mandated by the government to monitor the environmental, social, economic, political, and cultural developments at the regional level (PSA Resolution No. 10, Series of 2017). In this research, we are focusing on Region XII's non-metallics production quantity over the three-year period. According to the Organization for Economic Co-operation and Development (OECD), a non-metallic mineral commodity refers to a mineral usually having a dull luster, generally light-colored, transmits light, usually giving either colorless or light-colored streak from which a non-metallic element can be extracted for a profit such as a limestone, sand and gravel, clay, marble, silica, etc (OECD, 2001). The significance of this study to economic, social development, and technological is quite obvious (Pierdzioch C, Stadtmann G., 2013). Increasing demand for mineral goods and services result to economic expansion of developing countries. Hence, the future behavior of non-metallics production volume is vital for all agents of economy.

Region XII also known as SOCCSKSARGEN, is home to four provinces and one city composed of South Cotabato, Cotabato Province, Sultan Kudarat, Sarangani, and General Santos City. Using the non-metallics production data collected from different local government units (LGUs) from these provinces, consolidated by the Mines and Geosciences Bureau (MGB) Region XII, the main aim of the study is to develop, train, and evaluate linear and non-linear mathematical time series models. Forecasting is the prediction of future events. One way to forecast is via quantitative forecasting which is designed to predict the future value of a time series model with minimum data (Schnarrs, 1984). We also investigate seasonality in a de-trended macro-economic time series that often accounts to the movements of the production volume. Studies on the Philippines mineral production volume is quite scarce. Different analysts of market data mostly deal with forecasting mineral commodity prices instead of the production itself. One paper that forecast mineral production applies thermodynamics and Hubbert peak analysis in predicting mineral resources depletion (Valero, A. and Valero, A., 2010). They used the exergy analysis of minerals to forecast earth's physical stock of minerals and concluded that the peak of production might be reached before the end of the 21st century. Another study introduced the use of three-dimensional analytical model that can predict sanding onset from open hole wellbores (Al-Shaiibi, S., Al-Ajmi, A., Wahaibi, Y., 2013). It concluded that the develop model can be utilized as an approximation tool for perforated wells and predict onset pressure from wellbores.
As there are many different methods to forecast a phenomenon, a study revealed that no technique performs consistently well and better for all types of data (Lawrence, Edmundson, and O'Connor, 1983; Gardner and Dannenbring, 1980). Thus, more often than not, simple forecasting methods have better accuracy in forecasting than complex models (Armstrong, 1986; Lawrence, 1983).

II. Methodology

The Non-Metallics Production Dataset from the Department of Environment & Natural Resources – Mines and Geosciences Bureau Region XII is utilized and used with permission. The dataset contains only two variables, time and production quantity. The full dataset is shown below:

ID No.	Time	Production Quantity in '000 (cu.m.)
1	2017-01	172.0405
2	2017-02	179.775
3	2017-03	153.539
4	2017-04	133.3803
5	2017-05	153.783
6	2017-06	30.945
7	2017-07	149.444
8	2017-08	28.957
9	2017-09	13.6735
10	2017-10	164.7346
11	2017-11	208.5477
12	2017-12	79.9211
13	2018-01	180.7333
14	2018-02	225.7394
15	2018-03	329.4231
16	2018-04	477.6757
17	2018-05	334.8631
18	2018-06	289.3595
19	2018-07	250.1926
20	2018-08	195.1872
21	2018-09	201.5102
22	2018-10	310.0544
23	2018-11	251.4889
24	2018-12	265.797
25	2019-01	621.7637
26	2019-02	458.2238
27	2019-03	387.6178
28	2019-04	187.2189
29	2019-05	230.6134
30	2019-06	233.217
31	2019-07	332.2365
32	2019-08	232.0489
33	2019-09	129.2738
34	2019-10	348.5073
35	2019-11	234.2398
36	2019-12	280.9506

Table 1. Non-metallics production quantity in Region XII, Philippines.

The first variable deals with *time*. The second variable is the *production quantity in '000*. For example, the production volume in January 2017 is 172,040.50 cubic meters. In total, we are working on a 3-year time frame with a monthly non-metallics production data. Most of the non-metallics found and reported in Region XII are sand and gravel, limestone, boulders, and earth fills.

2.2 Training and Testing Set

For this study, the dataset is divided into two: 30 data points are used for training and 6 data points are taken for evaluation of the trained forecast algorithm. Thus, by forecasting the trained model on the unseen data, we can identify which model works and predicts best.

2.3 Implementation of the Different Forecasting Methods

As the aim of the study, we model the time series objects to collect and study past observations to develop an appropriate model to capture the structure of the series. The model is then used to create a forecast to generate future values. Different forecasting algorithms are compared for this study: Simple Average Method, Naïve Method, Seasonal Naïve Method, Regression Analysis, Stochastic Processes such as ARIMA Modeling, and Artificial Neutral Network.

2.3.1 Basic Forecasting Methods

These models are pretty straightforward and directly implemented in the dataset as benchmark methods against more complicated algorithms.

. Let $y_1, \dots, y_T$ be the historical data, T be the time series object and h be the forecast horizon.

- Simple Average Method – Uses an average of all past training data to forecast the next values. The model is presented as
$$\hat{y}_{T+h|T} = \frac{y_1, \dots, y_T}{T}. \#(1)$$
- Naïve – Uses the last period's actual value as a forecast. Thus,
$$\hat{y}_{T+h|T} = y_T. \#(2)$$
- Seasonal Naïve Method – Use for highly seasonal data. Setting each forecast to be equal to the last observed value from the same season of the year.
Let m be the seasonal period, and k be the number of complete years in forecast period prior to time $T + h$.
The model function is written as
$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}. \#(3)$$

2.3.2. Regression Analysis

Linear regression modeling is a powerful tool to explore causal relationships between events and predict future outcomes. That is why is it also widely used for forecasting. The simple linear regression model is given by the function
$$\hat{y} = b_1 X_1. \#(4)$$
Where b_1 is the least square estimate of the slope associated with X_1 (without loss of generality and for convenience, the intercept is always 0). Based on simple linear regression, we can develop a more complicated model such as the multiple linear regression which is given by the model function
$$\hat{y} = c_1 X_1 + c_2 X_2 \#(5)$$
Where c_1 *and* c_2 are the least-squares estimates of multiple linear regression parameters.
In this case, we use the Production Quantity as the response variable (y) and trend & seasonality $(X_1 + X_2)$ as the explanatory variables. The linear regression model is used to study the effect and impact of trend and seasonality on the production quantity.
Then, carry-out cross validation techniques on how well the model fit on the unseen data.

2.3.3 ARIMA Modeling

ARIMA Model is also known as the Box-Jenkins methodology where it is used to identify and diagnose time series data with the ARIMA model (Box and Jenkins, 1976). The ARIMA model is the linear combination of past values and errors in forecasting the value of a variable. Let y_T be the actual value, p and q be the autoregressive and moving average, respectively, ε_T be the random error at $T, and\ \emptyset_i\ and\ \emptyset_j$ are the coefficients. Hence, the forecast function is given by
$$y_T = \emptyset_0 + \emptyset_1 y_{T-1} + \emptyset_2 y_{T-2} + \dots + \emptyset_p y_{T-p} + \varepsilon_T - \theta_1 \varepsilon_{T-1} - \theta_2 \varepsilon_{2-1} - \dots - \theta_q \varepsilon_{T-q}. \#(6)$$
ARIMA Modeling has three major steps: *(1) Model Identification, (2) Model Estimation, and (3) Diagnostic Checking*. We first check for stationarity via plotting. The order of differencing is decided by using an R Statistical Software called the *forecast package* in R. The *auto.arima( )* function in R returns the best ARIMA model according to the performance metric of the AIC, AICc, or BIC value (Hyndman, 2008). The function conducts a search over a possible model within the order constraints provided. Thus, it tries all possible parameters for the given time series data and chooses a model that returns the lowest AIC, AICc, or BIC. The chosen model would be evaluated for further analysis.

2.3.4 Neutral Network in Time Series

The idea of ANN was first introduced in late 1943 by Walter Pitts and Warren S. McCulloch as a data processing unit for classification or prediction problems. A feed-forward neural network is consisting of an input layer, a hidden layer, and an output layer. The number of neurons in the input layer is the number of input attributes in the training dataset. The link between these neurons is called *weights.* The network is then followed by a hidden layer and an output layer. A neural network is also a deterministic model. Thus, we will feed our training data in different parameter settings of the network architecture to find the optimal *weights.*

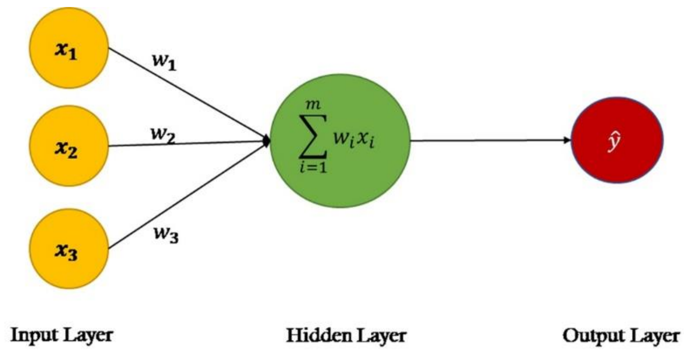

Figure 1. An example of a simple feed forward neural network adapted from "Single Neural Training" in Towards Data Science by Rojan Paleja, 2019, Retrieved May 22, 2020, from https://towardsdatascience.com/single-neuron-training-3fc7f84d67d.

The mathematical equation of the figure above is given by

$$y_i^{[l]} = w_i^{[l]^{Transpose}} x + b_i^{[l]}, \#(7)$$

where l be the lth layer of the network, i the ith unit in the lth layer, $w, b, and\ y$ are the weights, bias, and output, respectively. The weights and biases are *learned* from the data.
Time-series forecasts using neural network consists of a single hidden layer and lagged inputs for forecasting univariate time series. This is called a neural network autoregression or the NNAR model. The NNAR model is given by $NNAR(p, k)$ where p here is the lagged inputs and k nodes in the input layer. The best model is chosen according to the lowest AIC for the neural network model (Hyndman, 2008).

2.4 Performance Analysis

The success of an approximation algorithm depends on the performance, ease of implementation, and applicability.
It is important to note the kind of evaluation metric to use for measuring the forecasting effectiveness of the models. Every performance metric has its own advantages and disadvantages as elaborated more below (Adhikari & Agrawal, 2013).
In each definition, y_t is the actual value, f_t is the forecasted value, $e_t = y_t - f_t$ is the forecast error, and n is the test set size.

2.4.1 Mean Absolute Error (MAE)

The mean absolute error is defined by

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_t|. \#(8)$$

Using the mean absolute error as an accuracy indicator has its own advantages and disadvantages. MAE is usually used as a metric because it captures the magnitude of the overall error. A good MAE value should be as small as possible. The main property of MAE is that it measures the average absolute deviation of the forecast values from original ones. The disadvantage of using MAE and relying solely on MAE is that many positive and negative errors do not cancel out. It results in not penalizing extreme error values.

2.4.2. Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error is defined by

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_t}{y_t}\right| \ x\ 100. \#(9)$$

The Mean Absolute Percentage Error (MAPE) represents the percentage of average absolute error that occurred. MAPE is independent of the scale measurement, but is affected by data transformation and scaling. Thus, it does not penalize extreme deviations and opposite signed errors do not offset each other. Also, the disadvantage of using MAPE is that it does not show the direction of error.

2.4.3 Mean Squared Error (MSE)

The Mean Squared Error is defined by

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_t^2 . \#(10)$$

Mean Squared Error (MSE) is a famous forecast metric because it uses the average square deviation of forecasted values. As the opposite signed errors do not offset one another, MSE gives an overall idea of that occurred during forecasting. Compared to MAPE and MAE that do not penalize extreme value, MSE penalizes extreme errors occurred while forecasting. Hence, this forecast metric highlights the fact the total forecast error is affected by large individual errors, i.e., having one large error is much expensive than small errors combined. Also, MSE does not have any idea about the direction of the overall error. The disadvantage of MSE is that it is sensitive to scaling and data transformations. Thus, even if it is a good overall forecast measure, it is not easily interpretable as the other measures.

2.4.4 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is defined by

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_t^2} . \#(11)$$

The properties of RMSE is the same as MSE.

2.4.5 The Theil's U-Statistics

Theil's U-statistics is defined by [9]

$$U = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_t^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} f_t^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_t^2}} \# \qquad (12)$$

Theil's U-statistics is a normalized measure of the total forecast error. $0 \leq U \leq 1; U = 0$ means a perfect fit. As the same with MSE, this forecast metric is also affected by scaling and data transformation. A good Theil's U-Statistics value should be close to zero.

III. RESULTS AND DISCUSSION

Different forecasting algorithms are applied to the monthly production data set. Results are compared using different performance indicators such as the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Theil-U's Statistic. These forecast accuracy metrics calculated the forecasted value and the actual production quantity. For these metrics, the lower the value, the more accurate the forecast model. The computations were implemented using R programming language under Windows 10 Operating System with Intel Core i7-2.50 GHz machine and on 8 GB RAM.

3.1 Summary Statistics

We first look at the plots of the time series object. A total of 36 data points is plotted in a monthly production dataset below from the Year 2017-2019.

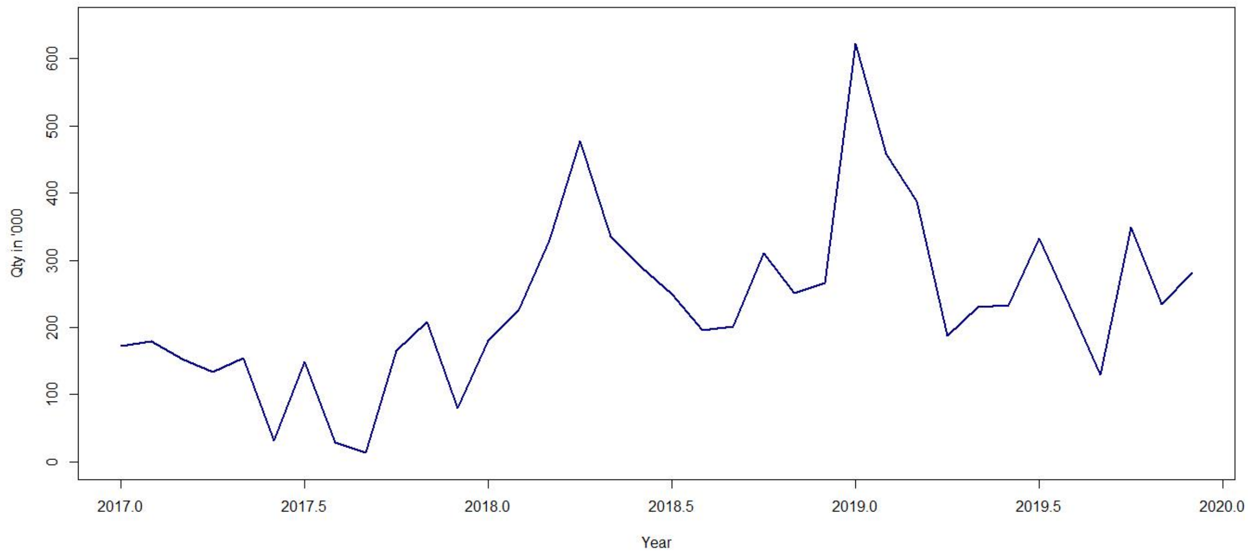

Figure 2. Monthly non-metallics production quantity over the years.

The non-metallics production series shown in Fig. 2 represents the monthly number of quantities from 2017-2019. The first 30 (i.e. January 2017 – June 2019) observations are used for training and the remaining 6 (i.e. July 2019-December 2019) for testing.

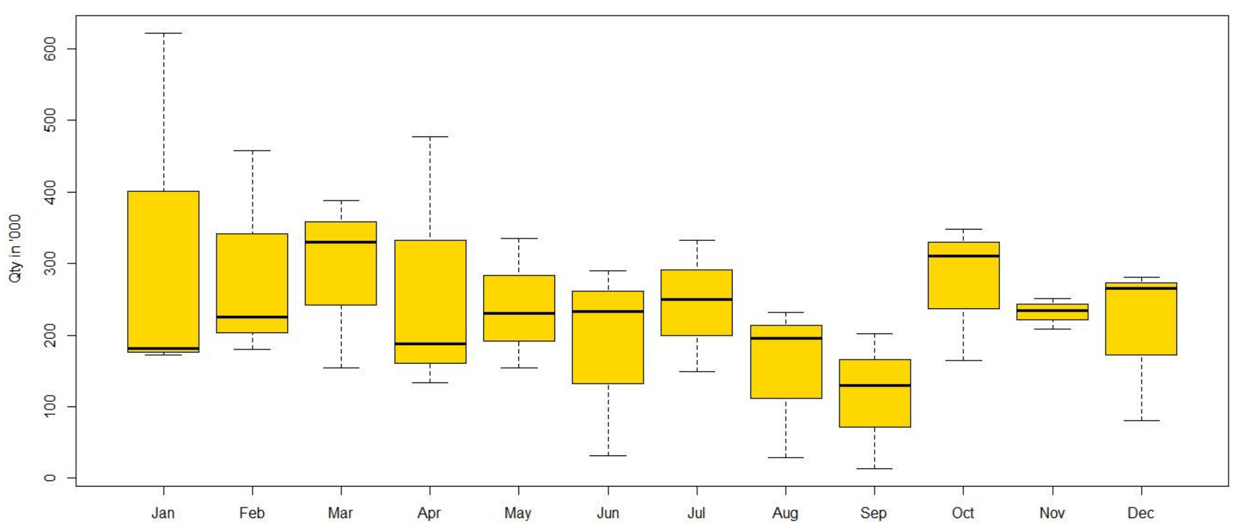

Figure 3. Seasonal distribution of the time series object.

From the boxplot, we can assume that seasonality has something to do with the Philippines' non-metallics production peaks. For March-May, which is considered summer and dry season in the Philippines, production on the minerals is high compared to the rainy season (June-September) every year.

On average, the produced quantities of non-metallic minerals in Region XII from the Year 2017-2019 is 234,910 cu.m. The lowest produced quantity is 13,670 cu.m. and the maximum produced quantity is 621,760 cu.m.

Method	MAE	RMSE	MAPE	Theil's U-Statistics
Average	55.250	76.058	24.485	0.595

Naïve	54.754	75.848	24.348	0.593
Seasonal Naïve	**50.498**	**55.130**	**23.155**	**0.306**
Univariate Regression Analysis (Production Quantity~Trend)	146.479	164.672	75.921	0.800
Multivariate Regression Analysis (Production Quantity~ Trend + Seasonality)	104.766	115.693	51.367	0.568
ARIMA Model -ARIMA(0,1,0)	54.123	72.754	23.430	0.600
Artificial Neural Network - NNAR(1,1,2)	115.776	143.193	63.240	0.718

Table 2. Results of different forecasting methods applied to the testing data set.

The result revealed that the forecasting performance of the Seasonal Naïve Algorithm is best in our experiments for the non-metallics production quantity dataset. The benchmark methods work well in this short time series consisting only of 30 training lagged inputs and tested on unseen data. Out of 7 different forecasting algorithms tested for the non-metallics production quantity in Region XII, Philippines, seasonal naïve outperformed all other methods when it comes to the given performance indicators, i.e., seasonal naïve showed the best forecast accuracy in terms of the different evaluation measures.

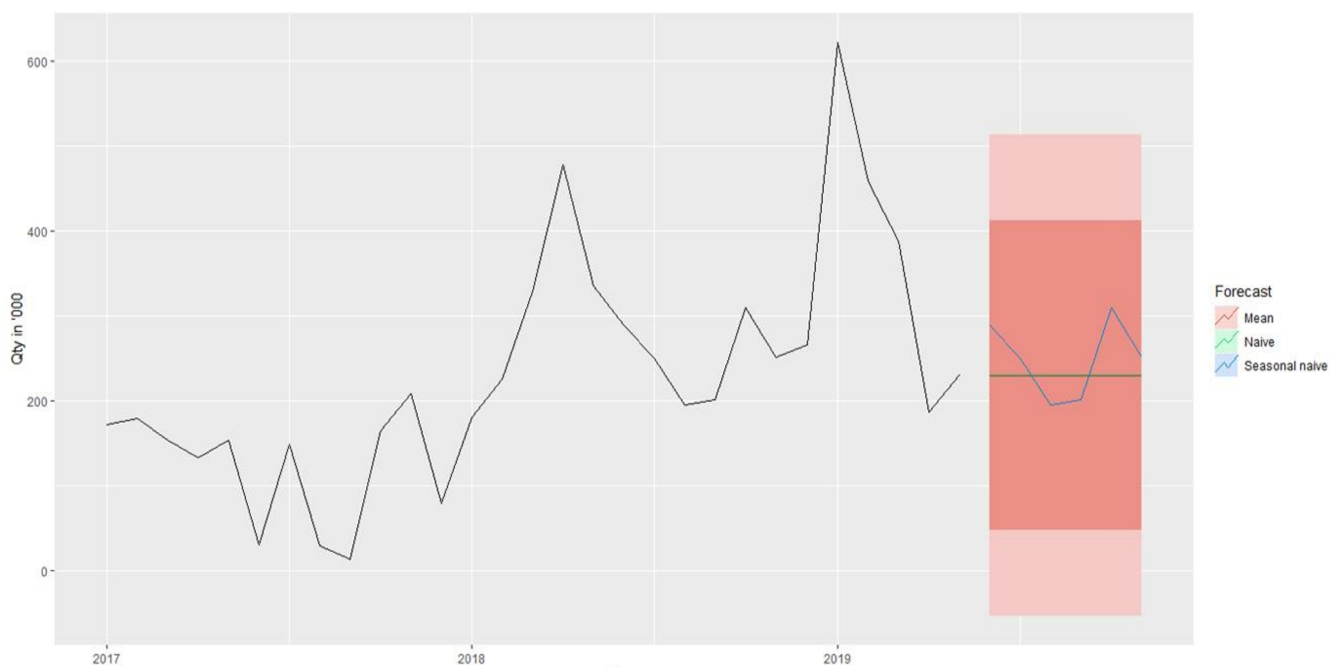

Figure 4. Comparative forecasting outcomes of benchmark models.
It is apparent that mean and naïve methods performed poorly compared to seasonal naïve.

3.2 Regression Analysis

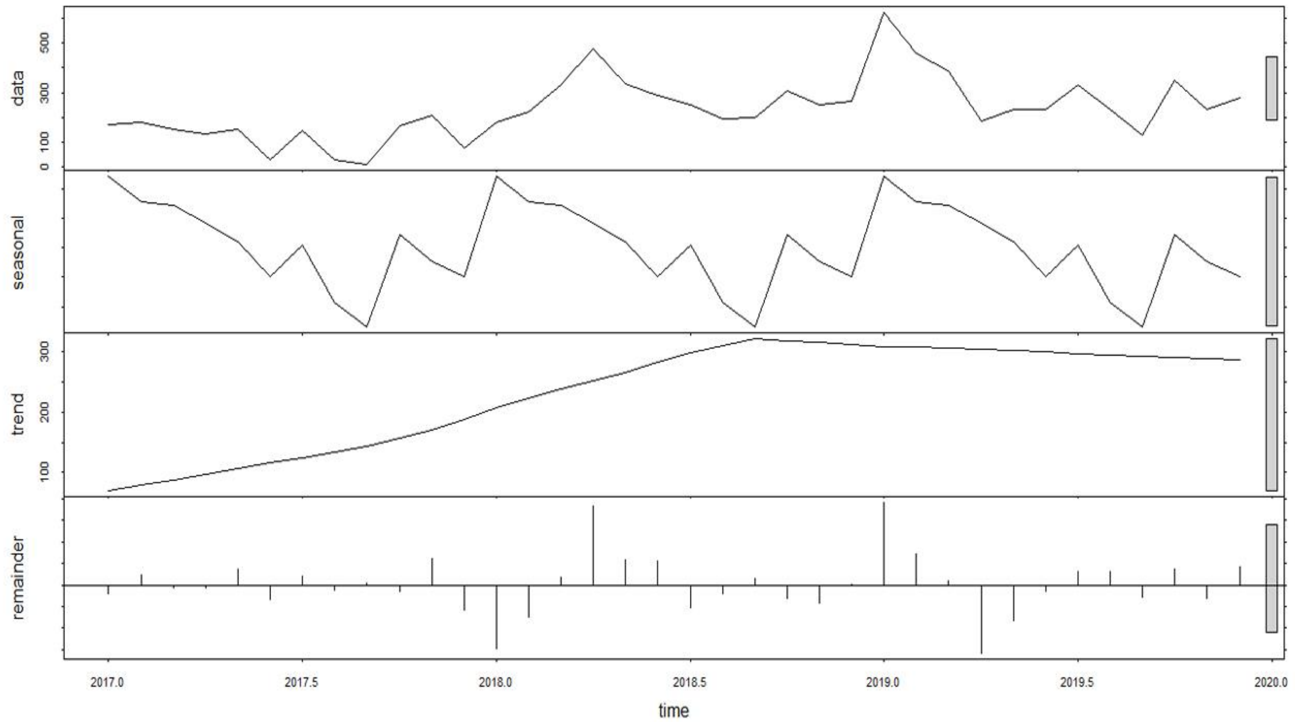

Figure 5. Seasonal and trend decomposition using Loess (STL Decomposition).

We get a good sense of how the time series behaves from the graph above. The figure is a decomposition tool that separates trend, seasonality, and random noise individually. Regression Analysis is then used to see if seasonality and trend truly affect production quantity. From this chart, we can see that seasonality is strong from 2018 to 2019.

```
Call
tslm(formula = sr ~ trend)

Residuals:
     Min       1Q   Median       3Q      Max
-167.597  -82.738    5.482   42.413  295.782

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   85.700     42.204   2.031 0.052249 .
trend          9.611      2.457   3.911 0.000559 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 110.7 on 27 degrees of freedom
Multiple R-squared:  0.3617,    Adjusted R-squared:  0.3381
F-statistic:  15.3 on 1 and 27 DF,  p-value: 0.0005591
```

Table 3. Fit statistics for univariate linear regression model via R Statistical Software (trend in the time series object affects the production quantity).

Linear regression did a good job of picking up the trend, that is, the explanatory factor is significant. The estimated model is defined as

$$\hat{y} = b_1 X_1 = 9.611 trend.$$

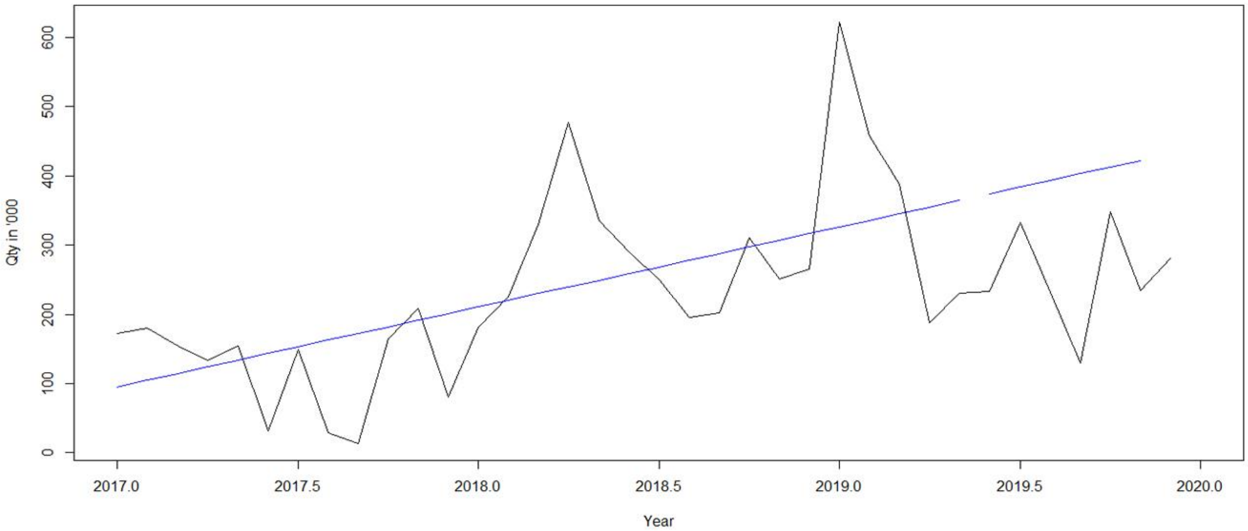

Figure 6. Scatterplot between fitted and actual values on predicting outcomes via simple linear regression analysis.

However, we will also test the significance of seasonality in the time series.

```
Call:
tslm(formula = sr ~ trend + season)

Residuals:
    Min      1Q  Median      3Q     Max
-200.10  -32.32  -10.24   35.26  211.58

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  193.513     71.005   2.725 0.014973 *
trend         10.102      2.484   4.067 0.000897 ***
season2      -47.036     89.464  -0.526 0.606272
season3      -54.858     89.567  -0.612 0.548831
season4      -89.062     89.739  -0.992 0.335751
season5     -125.503     89.980  -1.395 0.182142
season6     -154.591    100.016  -1.546 0.141733
season7     -125.028     99.985  -1.250 0.229109
season8     -222.876    100.016  -2.228 0.040545 *
season9     -237.459    100.108  -2.372 0.030572 *
season10    -117.759    100.262  -1.175 0.257377
season11    -135.237    100.477  -1.346 0.197081
season12    -202.499    100.753  -2.010 0.061621 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109.5 on 16 degrees of freedom
```

```
Multiple R-squared:  0.6298,    Adjusted R-squared:  0.3521
F-statistic: 2.268 on 12 and 16 DF,  p-value: 0.06376
```

Table 4. Fit statistics for multivariate regression model via R Statistical Software (trend and seasonality affect the production quantity).

It did get a better R^2 at 0.6298 compared to the univariate model but adding seasonality just worsened the model. All the seasons are not significant and thus should not be added to the model. However, when it comes to predicting future values, the multivariate model is better at forecasting. It gave a better forecast accuracy compared to the univariate model as shown below.

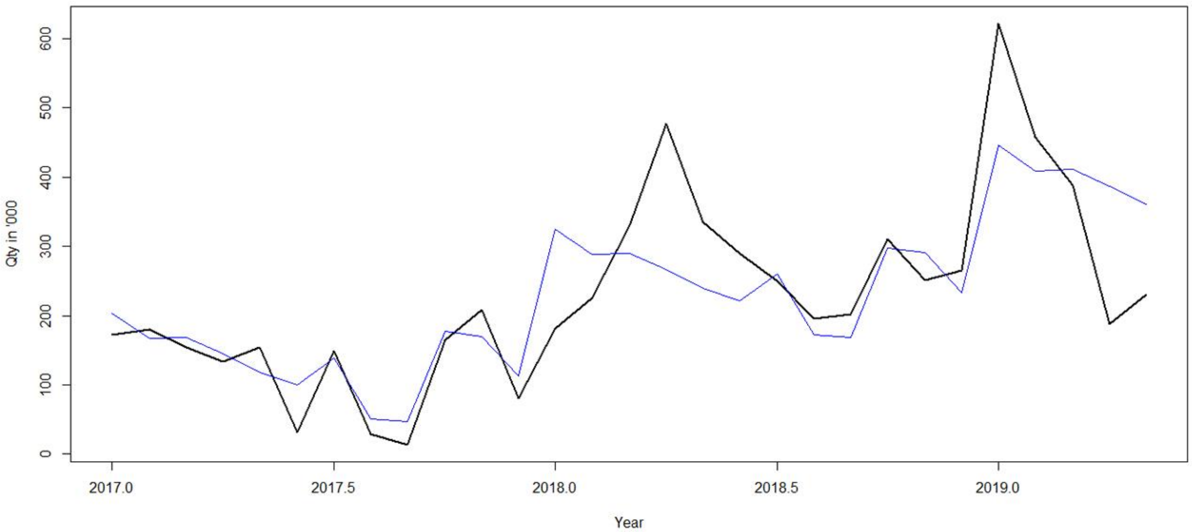

Figure 7. Results of observed and actual values fitted via multivariate regression analysis.

3.3 ARIMA Modeling

The forecast package in R has an *auto.arima()* function that finds the optimal and estimated ARIMA model that chooses the least AIC score.

```
ARIMA(0,1,0)
Series: sr

ARIMA(0,1,0)


sigma^2 estimated as 13379:  log likelihood=-172.75
AIC=347.5   AICc=347.66   BIC=348.83
```

Table 5. Result of auto.arima via R Statistical Software.

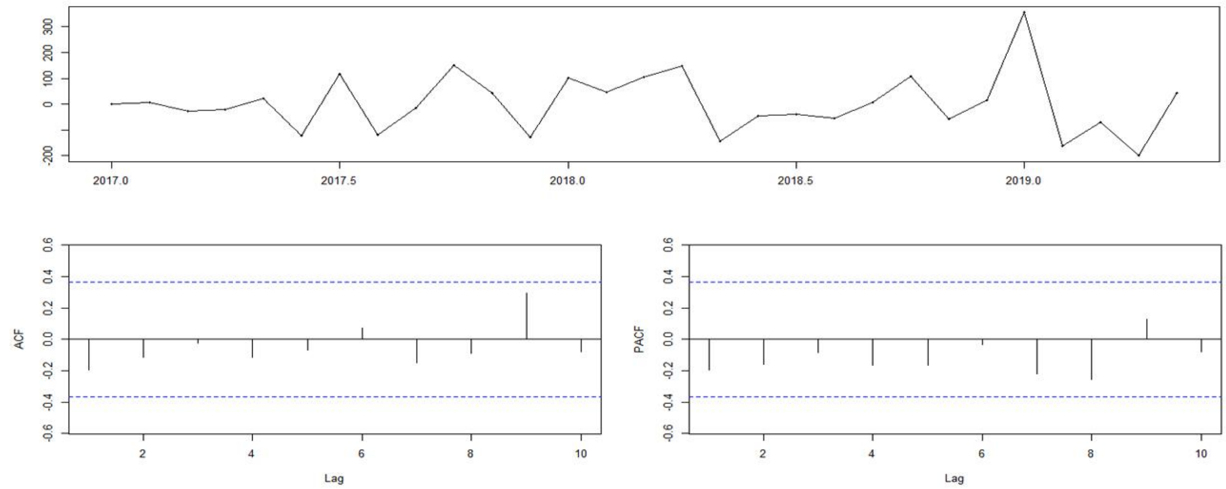

Figure 8. ACF and PACF plots of the time series model.

$ARIMA(0,1,0)$ has been chosen as the optimal model configuration for the time series. Thus, the time series does not contain the AR component, i.e., it showed a very slight dependence as seen in the auto-correlation and partial auto-correlation plots above. It also does not contain the moving average. It only contains the I-component, meaning the best model has stationarity in it as expected from the time series plot.

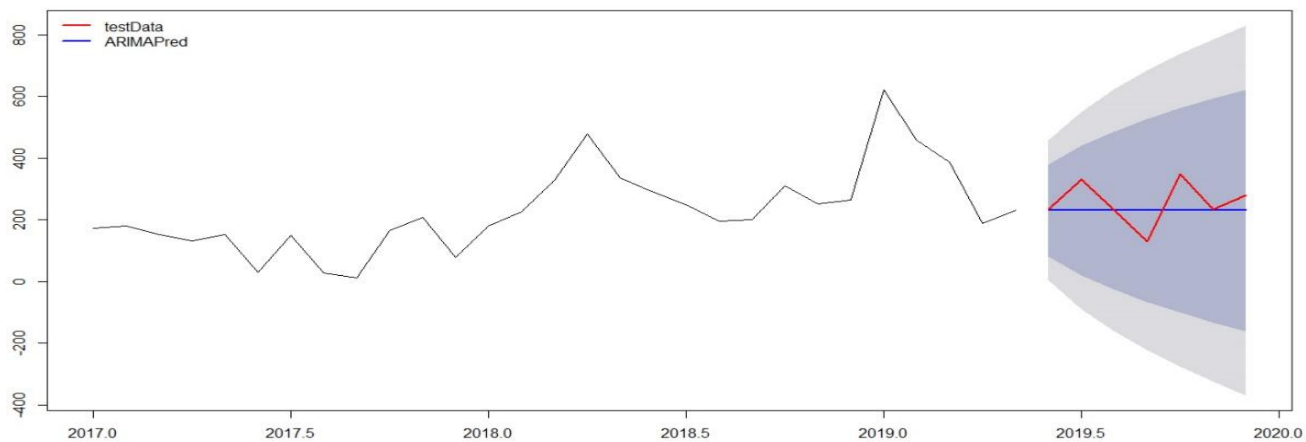

Figure 9. ARIMA modeling prediction plot.

3.4 Artificial Neural Network (ANN) Autoregression

```
Model:  NNAR(1,1,2)[12]

Average of 20 networks, each of which is
a 2-2-1 network with 9 weights
options were - linear output units

sigma^2 estimated as 6497
```

Table 6. NNAR results via R statistical software.

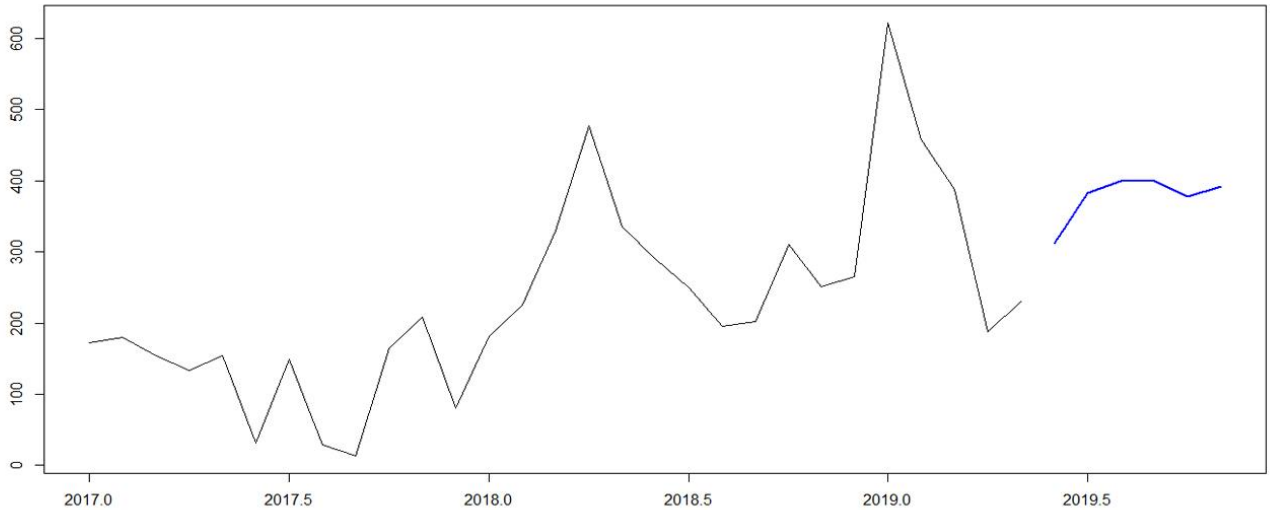

Figure 10. NNAR model prediction plot.

The model selection for the NNAR model is chosen via the lowest AIC. Thus, the NNAR(1,1,2) is chosen as the optimal model for this time series, that is, 1 lagged input is used as a predictor, and 1 hidden layer containing 2 neurons are used.

IV. CONCLUSION

The results indicate that different time series models can improve production forecast outcomes. Hence, out of all the models tested such as the average method, naïve method, seasonal naïve method, simple linear regression analysis, multiple regression analysis, ARIMA modeling, and Artificial Neural Network, the seasonal naïve method outperformed all other models used for this paper.

The outcome of this research can be used as a benchmark paper in employing more exogenous variables that can account for the non-metallics production quantity in Region XII. Furthermore, this study can be easily replicated by other researchers. But since this study was retrospective, it imposed no controls over the data and the data itself is representative of the actual operations.

This study is important because it provides an opportunity to create better policy frameworks that lead to larger revenues that help sustain projects of the government.

Appendixes, if needed, appear before the acknowledgment.

ACKNOWLEDGMENT

The author is grateful to the Department of Science and Technology – Science Education Institute (DOST-SEI) for granting the opportunity to research. At the same time, she is thankful to her research institution – Mines and Geosciences Bureau Region XII (MGB XII) for providing all the encouragement and data needed to push through the research. To everyone who has been part of this fruitful journey, from the staff to her family, a warm thank you.

REFERENCES

Adhikari, R. & Agrawal, R. (2013). An Introductory Study on Time Series Modeling and Forecasting. *LAP Lambert Academic Publishing, arXiv:1302.6613*.

Armstrong, J.S. (1986). Research on forecasting: A quarter-century review, 1960-1984. *Interface, 16,* 77-89.

Box, G.E.P. and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control (San Francisco: Holden-Day).

Gardner, E.S., and Dannenbring, D.J. (1980). Forecasting with exponential smoothing; Some guidelines for model statistics. *Decision Science*, *11*, 370-383.

Hyndman, RJ and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R, *Journal of Statistical Software*, 26(3).

Lawrence, M.J., Edmundson, R.H., and O'Connor, M.J. (1983). An exploration of some practical issues in the use of quantitative forecasting models, *Journal of Forecasting, 2,* 169-179.

McCulloch, W.S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biophysics.* Vol 5, pp. 115-133.

OECD (Organization for Economic Co-operation Development). (2001). Non-metallics Mineral Reserves. Retrieved May 19, 2020, from https://stats.oecd.org/glossary/detail.asp?ID=1815

Pierdzioch C., Rulke JC., Standtmann, G. (2013). Forecasting metal prices: do forecasters herd? J Bank Finance. 37(1): 159-8.

PSA (Philippine Statistics Authority). (2017). Approving and Adopting the Core Regional Indicators. Retrieved from https://psa.gov.ph/sites/default/files/PSA%20BOARD%20RESO%2010%2C%20S%202017_F.pdf

Schnarrs, S. (1984). Situational factors affecting forecasting accuracy. *Journal of Marketing Research,* 21-290-297.

Valero, A. and Valero, A. (2010). Physical genomics: Combining the exergy and Hubbert peak analysis for predicting mineral resources depletion. Journal of Resources, Conservation, and Recycling. Vol. 54, Issue 12, P 1074-1083.

AUTHOR

Sheila Mae D. Golingay – Department of Environment and Natural Resources – Mines and Geosciences Bureau XII