Time Series Analysis of Corn Production in the United States in 1900 to 2020

Gilbert Aurelio S. (23101910047) , Genta Ananda P.K. (23101910001) , and Dennis Jonathan (23101910027).

Business Mathematics 2019
School of Science, Technology, Engineering and Mathematics, Universitas Prasetiya Mulya, BSD City Kavling Edutown I.1, Jl. BSD Raya Utama, BSD City 15339, Kabupaten Tangerang, Banten, Indonesia

gilbert.sachio@student.pmsbe.ac.id, genta.kharisma@student.pmsbe.ac.id, dennis.jonathan@student.pmsbe.ac.id

Abstract. Agriculture is one of the challenges that must be faced in the future. One of the crops that people often need most is corn. According to Britanica, there are a lot of uses of corn, such as food for the people, ingredients for cooking, feed for livestocks, industrial uses such as making fructose-syrups and alcohols, and in the coming age of renewable energy, as a component for biofuel. This study aims to predict how much corn might be produced in a country given the knowledge of historical numbers and are there any interesting patterns which can be seen in those numbers. To predict how much corn might be produced, we will use autocorrelation and autoregressive moving average. The result showed that Corn production in the United States from 1900 to 2020 has a tendency to continuously increase overtime, thus creating a positive trend. But the data has random volatility lag without any seasonality, therefore it is possible to attempt to model the data using Holt's Exponential Smoothing and ARMA. but, it is better to utilize Holt's Exponential Smoothing since it performs better in almost every model performance metric.

1. Introduction

Agriculture is one of the challenges humankind is facing in this rapidly changing world. As of July 2021, as the human population is slowly edging towards the eight billion mark, there are more and more people who need to be fed, which leads to the increase of demands for some food sources such as crops. Feeding people is not the only issue as most crops also have alternative uses, thus humanity is in a constant loop of always needing more and more.

One of the crops that people often need most is corn. Corn, scientifically referred to as *Zea Mays, is* undeniably woven to the roots of our society. According to Britanica, there are a lot of uses of corn, such as food for the people, ingredients for cooking, feed for livestocks, industrial uses such as making fructose-syrups and alcohols, and in the coming age of renewable energy, as a component for biofuel [1,2]. This multi-use property of corn should be something to be taken into consideration as this might also bring some positive and negative effects. Fromme (2019) stated the demand for plants such as corn increased as there is an increase in demand due to the swiftly growing population and also the increase in biofuel production [3]. Capehart (2019) found that United States' farmers planted more corn than any other crops for 2019, accounting for roughly 91.7 million acres of land [4].

This brings right into the problems, is there a way to predict how much corn might be produced in a country given the knowledge of historical numbers and are there any interesting patterns which can be seen in those numbers. The authors have decided to analyse the production of corn in the United States of America, in particular the national production of corn grain measured in Bushels abbreviated as Bu (a unit of volume), using time-series methods. The dataset will be taken from surveys conducted by the United States Department of Agriculture which is available on their website. The time period will be limited between 1900 and 2020.

2. Literature Review

2.1. Time Series Data

Montgomery (2016) defined time series is a time-oriented or chronological sequence of observations on a variable of interest [5]. In a time series data, time is often the independent variable or described by Brownlee (2016) as "both a constraint and a structure that provides a source of additional information" [6]. The goal is usually to make a forecast for the future.

2.2. Stationarity, Differencing, and the Augmented Dickey-Fuller Test

Hyndman (2018) argued that a stationary time series is one whose properties do not depend on the time at which the series is observed [7]. Should a data exhibit patterns such as trends and seasonality, it is assumed to be not stationary.

There are ways of creating stationary data out of non-stationary data; Hyndman proposed the use of a method called differencing to remove changes between data. A first order differencing can be defined mathematically as:

$$y_t' = y_t - y_{t-1} \tag{1}$$

where $y_t^{'}$ is the value of present differenced observation, y_t is the value of the present observation, y_{t-1} the value of the previous observation.

Brownlee (2016) stated that one of the ways one can test stationarity of data is to use the Augmented Dickey-Fuller Test or also referred to as the Unit Root Test [8].

2.3. Autocorrelation function and Partial Autocorrelation Function

Smith (2021) defined autocorrelation as a scientific representation of the similarity between a given time series and a lagged version of itself over time [8]. Autocorrelation function (ACF) is the correlation between the values of a time series that are equal to the time lag of 0, 1, 2 periods or more. Meanwhile, Tinungki (2018) defined Partial Autocorrelation Function (PACF) as a function to measure the level of closeness between Zt and Zt+k if the effect of time lag is 1, 2, . . ., k-1 is considered separately [10].

2.4. Autoregressive Moving Average (ARMA)

Mitrović (2010) defined autoregression as a technique where a predictor estimates the value using a linear combination of the previous values [11]. Hyndman (2018) stated that the moving average is a regression which utilises the past forecast errors to form a prediction [6]. The order from MA can be seen from the ACF figure, while one can get the order for AR from the PACF figure. ARMA itself is a combination of AR and MA. It is written mathematically as:

$$y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q} \tag{2}$$

where p is the lag order for autoregression, $q$is the lag for moving average, y_{t-p} is he observed value at given lag (p). ϵ_{t-q} is the residual value at given lag (q), $\phi_p$is The parameter for the autoregression at given lag (p), θ_q is the parameter for the moving average at given lag (q).

2.5. Holt Exponential Smoothing (Double Exponential Smoothing)

According to Habsari (2020), the double exponential smoothing method is an exponential smoothing method which considers the presence of a trend pattern in the data [12]. The smoothing method is defined mathematically as :

The exponentially smoothed series or current level estimate :

$$L_t = \alpha y_t + (1 - \alpha)[L_{t-1} + b_{t-1}] \tag{3}$$

The trend estimate where b_t is defined as:

$$b_t = \beta[L_t - L_{t-1}] + (1 - \beta)b_{t-1} \tag{4}$$

Forecast m-periods into the future :

$$F_{t+m} = L_t + mb_t \tag{5}$$

where L_t is the level at time t, α is the weight for the level, B_t is the trend at time t, β is the weight for the trend, y_t is the data value at time t, and F_{t+m} is the m-step-ahead forecast, at time t.

2.6. Mean Absolute Percentage Error (MAPE)

According to Kim (2016), Mean Absolute Percentage Error (MAPE) is the average of absolute percentage errors (APE) [13]. Let At and Ft denote the actual and forecast values at data point t, respectively and n as the number of data poin, then, MAPE is defined mathematically as:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100 \quad (6)$$

2.7. Root Means Square Error (RMSE)

Glen (2021) defined Root Mean Square Error (RMSE) as the standard deviation of the residuals (prediction errors) or simply how far residuals are from the regression line data points[14].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2} \quad (7)$$

Where *Oi* are the observations, *Si* predicted values of a variable, and *n* the number of observations available for analysis.

2.8. Akaike Information Criterion and Bayesian Information Criterion

Akaike Information Criterion or often abbreviated as AIC and also Bayesian Information Criterion or Schwarz Information Criterion, often abbreviated as BIC, are metrics which assess the model's performance.

According to Gudjarati (2009), AIC and BIC impose higher penalties to models that say adjusted R-squared. BIC itself also imposes a harsher penalty compared to AIC since it accounts for the number of input rows in the model. The formula for both AIC and BIC are given below:

$$AIC = -2log(L) + 2p \quad (8)$$

$$BIC = -2log(L) + p\,ln(n) \quad (9)$$

where $L$is the likelihood function, $p$is the number of parameters and $n$is the number of rows which the model used as input[15].

2.9. Previous Studies

Prior to this research, Askar Choudhury and James Jones (2014) had made predictions about crop yields in several regions of Ghana from 1992 - 2008 by comparing the use of smoothing and ARMA. From this study, it was found that the use of the smoothing or ARMA method was determined by the dynamics of the data. They found that ARMA is better than a smoothing model from the best value R-squared in the ARMA model, but the value of AIC isn't the best. In conclusion, the ARMA model ignores variance data between the lag of years over time. [16]

This study will also look for the best method to predict, but it is preceded by looking for trend and seasonality so that it is more effective in choosing smoothing and ARMA types.

3. Methodology

3.1. Data for Analysis

As mentioned in the introduction, the dataset the authors used is available from the United States Department of Agriculture under the category "CORN, GRAIN - PRODUCTION, MEASURED IN BU". The data ranged from the year 1866 up to 2020, but for this analysis, the author used data starting from 1900 up to 2020. The original dataset contained some columns that were unnecessary for the analysis, thus the authors decided to not include them and focused on the column Value instead.

3.2. Data Exploration

The analysis was conducted using the programming language *Python* using the kernel available on the website Deepnote. The authors utilized some libraries to help us in modelling the dataset. For time-series analysis, the authors used the library Statsmodels.

Before conducting the time-series analysis, there is a need to determine whether the given dataset is stationary or not. A method of determining the stationarity is using the Augmented Dickey-Fuller test and using the p-value result to determine whether the data is stationary or not relative to a certain significant level or α (the authors used $\alpha = 0.05$). Should the data be not stationary, the author will perform differencing and repeat the Augmented Dickey-Fuller test. After the Augmented Dickey-Fuller test, the author decomposed the data to see whether there were trend and seasonal patterns in the data.

The authors started analysing the data using Holt Double Exponential Smoothing to model the pre-differenced data. After using Holt Double Exponential Smoothing, the author followed the analysis by previewing the ACF and PACF of the differenced data in order to see some estimates of parameter AR and MA to use the ARMA time-series model. After that, the author makes the function for finding the best parameter ARMA and plotting between actual data and predicted data.

The modelling section was concluded by comparing the given result from Holt Double Exponential Smoothing and the ARMA model using the best parameters. This comparison included checking some error metrics, error distribution, and forecasting the next ten years.

4. Result and Analysis

4.1. Stationarity Testing

The data which the author had gathered from the U.S. Bureau of Labor Statistics was tested using the Augmented Dickey-Fuller Test. The result is the p-value of 0.999, thus the null hypothesis was rejected, confirming the presence of stationarity in the dataset.

As mentioned in Chapter 2, Hyndman proposed that to get rid of the stationarity problem is to perform differencing on the data. The author performed the first order differencing and tested this new data using the Augmented Dickey-Fuller Test and received a p-value of $4.408*10^{-14}$ which is roughly equivalent to 0, therefore this differenced data can be considered as non-stationary.

4.2. Testing for Seasonality

A seasonal decomposition was conducted to see whether there is indeed the presence of seasonal influence on the actual dataset. The result for the seasonal decomposition can be seen on **Figure 4.2.1**. It is noticeable that the decomposition did not find any presence of seasonality, rather the movement of the data is the result of the trend itself.

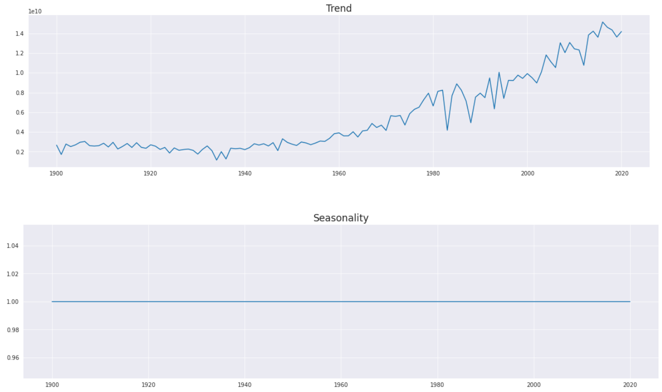


Figure 4.2.1. Seasonal Decomposition Result

4.3. Modelling Process

Due to the presence of a trend factor, Holt's Exponential Smoothing was a candidate model for the dataset. After performing Holt's Exponential Smoothing on the pre-differenced dataset, the parameter for the given data were $\propto = 0.2407$ and $\beta = 0.0001$. It could be seen that there is indeed the presence of a trend factor, but the impact of it is not as significant. The result of Holt's Exponential Smoothing can be seen on **Figure 4.3.1.**

Since the data after the differencing process exhibited no stationary problem, an autoregressive moving average (ARMA) model was also considered to be a candidate model. Determining the order of the autoregression as well as the moving average required both the autocorrelation function (ACF) plot and also the partial autocorrelation function (PACF) plot.

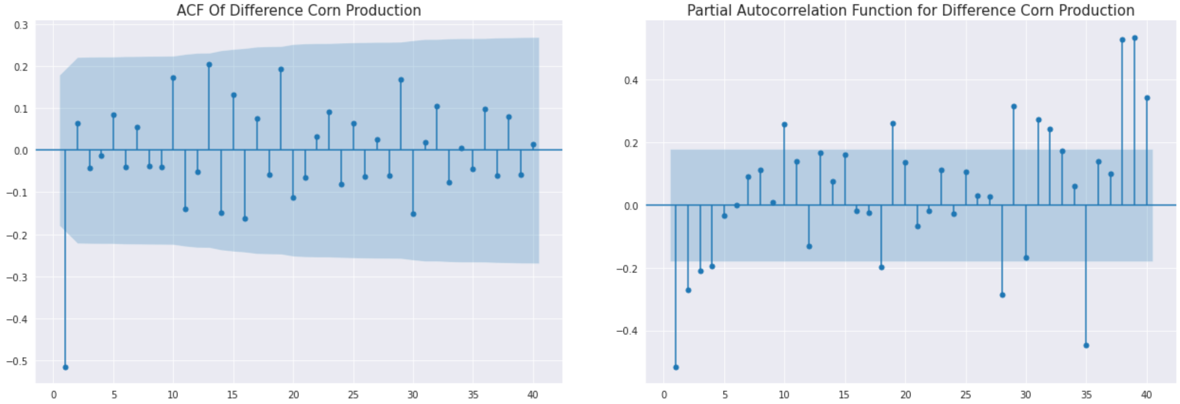


Figure 4.3.1. ACF and PACF plots for the Differenced Data

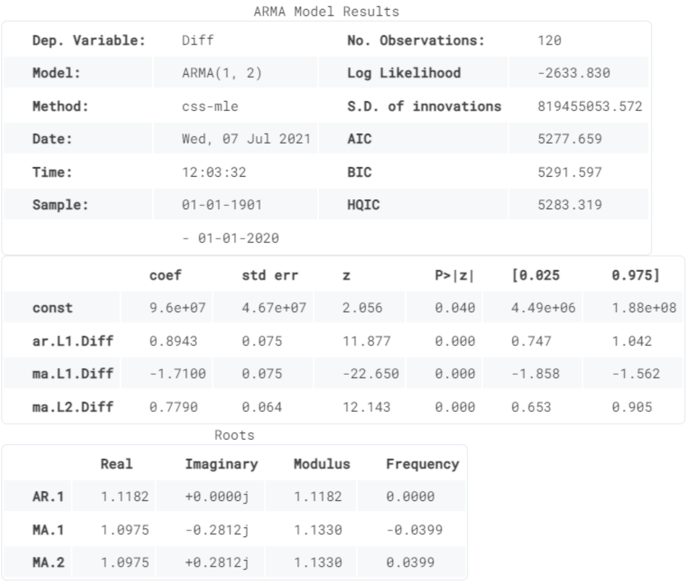
The autocorrelation function plot and the partial autocorrelation function plot indicated a possibility of utilizing ARMA(1,1). Another possibility of finding the right combination is doing gridsearch on several candidates. After conducting the gridsearch using 100 possible candidate ARMA models and evaluating them based on the log likelihood ratio test, the final result was the model ARMA(1,2) with the equation for the differenced data:

$$\widehat{z_t} = 0.8943y_{t-1} + \epsilon_t - 1.71\epsilon_{t-1} - 0.779\epsilon_{t-2} + 9.6 \times 10^7 \quad (10)$$

In order to convert the differenced prediction into the pre-differenced prediction, one could add the differenced prediction to the actual value at the previous lag, thus creating the formula:

$$\widehat{y_t} = y_{t-1} + 0.8943y_{t-1} + \epsilon_t - 1.71\epsilon_{t-1} - 0.779\epsilon_{t-2} + 9.6 \times 10^7 \quad (11)$$

The summary for the ARMA(1,2) can be seen below:



ARMA Model Results

Dep. Variable:	Diff	**No. Observations:**	120
Model:	ARMA(1, 2)	**Log Likelihood**	-2633.830
Method:	css-mle	**S.D. of innovations**	819455053.572
Date:	Wed, 07 Jul 2021	**AIC**	5277.659
Time:	12:03:32	**BIC**	5291.597
Sample:	01-01-1901	**HQIC**	5283.319
	- 01-01-2020		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	9.6e+07	4.67e+07	2.056	0.040	4.49e+06	1.88e+08
ar.L1.Diff	0.8943	0.075	11.877	0.000	0.747	1.042
ma.L1.Diff	-1.7100	0.075	-22.650	0.000	-1.858	-1.562
ma.L2.Diff	0.7790	0.064	12.143	0.000	0.653	0.905

Roots

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	1.1182	+0.0000j	1.1182	0.0000
MA.1	1.0975	-0.2812j	1.1330	-0.0399
MA.2	1.0975	+0.2812j	1.1330	0.0399

Figure 4.3.2. Summary of ARMA(1,2)

In Figure **4.3.2,** we can see that all parameters in the ARMA(1,2) result are indeed significant. The roots also tell us that every lags are indeed stationary processes.

4.4. Comparison of Holt's Exponential Smoothing and ARMA(1,2)

From **Figure 4.4.1**, Holt's Exponential Smoothing and ARMA(1,2) look similar in both how they fit the regressor and also how they forecast. But in detail, ARMA (1,2) has a tendency more than Holt's Exponential Smoothing. Both models also captured the trend in the data set magnificently, but are not very good in following the rapid changes in the data . Underfitting model is more preferable because it shows the mean of the randomness and the trend more.

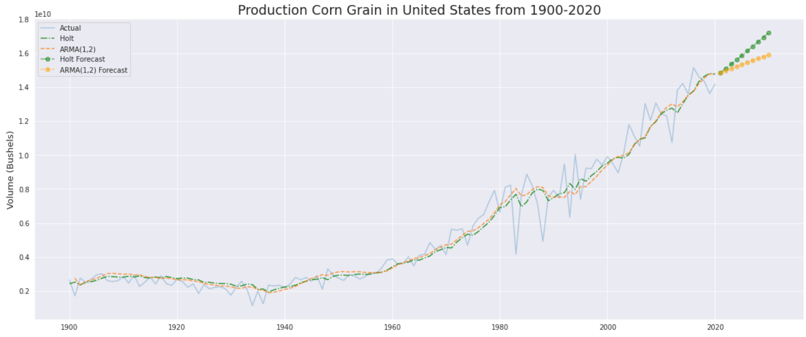


Figure 4.4.1. The Result of the Models and Forecasting the Next 10 Entries

From **Figure 4.4.2**, one can observe the errors for both candidate models. There is no clear pattern on any of the residuals for every input value for both models. From the Q-Q plot and the kernel density estimate plot, it can be deduced that the residuals are not normally distributed since the residuals fail to follow the Q-Q plot reference line, have a slight skew (positive skew for Holt's Exponential Smoothing and negative skew for ARMA(1,2), as well as having a rather large kurtosis for both models. It can also be seen that there are some outliers for higher quantiles in the ARMA(1,2) residuals and outliers on the lower quantiles for Holt's Exponential Smoothing.

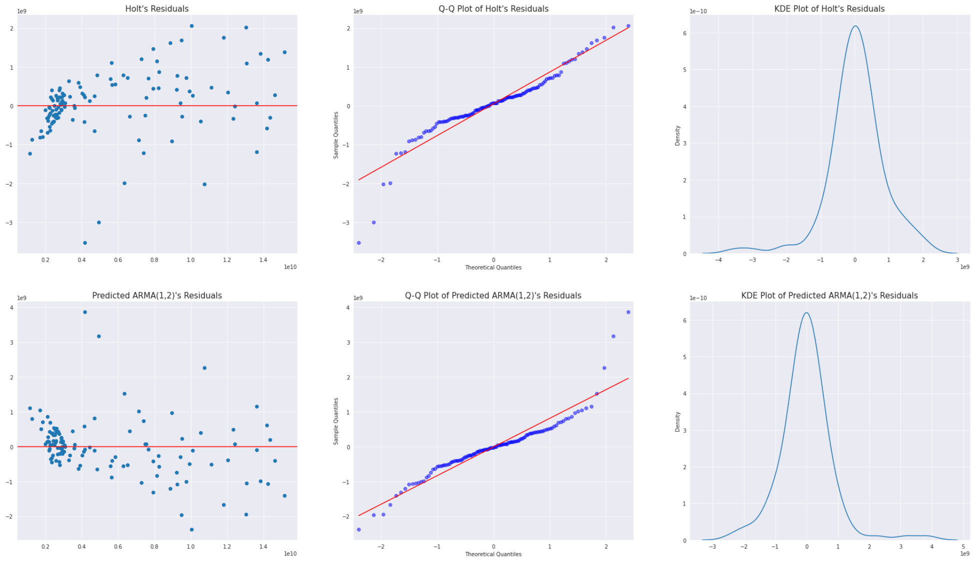


Figure 4.4.2. Residual plot, Q-Q plot, and Kernel Density Estimate plot residual from Holt's Exponential Smoothing and ARMA (1,2)

Comparing the models from its model performance metrics can also be beneficial. Some metrics for both models are available on Table 4.4.1. Holt's Exponential Smoothing has a lower root mean squared error score compared to ARMA(1,2) which indicates the model's capability to be closer towards the actual corn production values. ARMA(1,2) performs better in mean absolute percentage error, therefore when combined with the knowledge of the root mean squared error metric, one could think that ARMA(1,2) generally has bigger errors which are amplified by squaring the errors rather than getting the absolute value. The Akaike Information Criterion and Bayesian Information Criterion showed the upper hand for Holt's Exponential Smoothing since it is generally accepted that ,the lower those metrics are, the better the model.

Table 4.4.1 Comparison of Model Performance Metrics

	RMSE	**MAPE**	**AIC**	**BIC**
Holt's Exponential Smoothing	818,114,044	12.51855%	4974.4487	4985.6311
ARMA (1,2)	822,131,658	12.36349%	5277.6591	5294.6342

5. Conclusion

Corn production in the United States from 1900 to 2020 has a tendency to continuously increase overtime, thus creating a positive trend. But the data has random volatility lag without any seasonality, therefore it is possible to attempt to model the data using Holt's Exponential Smoothing and ARMA.

The result of the Holt's Exponential Smoothing are a model with parameters of $\propto = 0.2407$ and $\beta = 0.0001$. The best model parameter for ARMA is (1,2) with the equation:

$$\widehat{z}_t = 0.8943y_{t-1} + \epsilon_t - 1.71\epsilon_{t-1} - 0.779\epsilon_{t-2} + 9.6 \times 10^7 \qquad (10)$$

Both models are able to capture the given input very well, but for this particular data, it is better to utilize Holt's Exponential Smoothing since it performs better in almost every model performance metric. Holt's Exponential Smoothing underfits therefore the model shows the mean of randomness. The underfitting model also focuses more on the trend in forecasting.

6. References

[1] The Editors of Encyclopaedia Britannica. (n.d.). *Corn | History, Cultivation, Uses, & Description*. Encyclopedia Britannica. Retrieved July 6, 2021, from https://www.britannica.com/plant/corn-plant

[2] Foley, J. (2013, March 5). *It's Time to Rethink America's Corn System*. Scientific American. https://www.scientificamerican.com/article/time-to-rethink-corn/

[3] Fromme, D. D., Spivey, T. A., & Grichar, W. J. (2019). Agronomic Response of Corn (Zea maysL.) Hybrids to Plant Populations. *International Journal of Agronomy*, *2019*, 1–8. https://www.hindawi.com/journals/ija/2019/3589768/

[4] Capehart, T., & Proper, S. (2019, July 29). *Corn is America's Largest Crop in 2019*. USDA. https://www.usda.gov/media/blog/2019/07/29/corn-americas-largest-crop-2019

[5] Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction to Time Series Analysis and Forecasting (Wiley Series in Probability and Statistics)* (2nd ed.). Wiley-Interscience.

[6] Brownlee, J. (2020a, August 14). *How to Check if Time Series Data is Stationary with Python*. Machine Learning Mastery. https://machinelearningmastery.com/time-series-data-stationary-python/

[7] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). OTexts.

[8] Brownlee, J. (2020a, August 14). *How to Check if Time Series Data is Stationary with Python*. Machine Learning Mastery. https://machinelearningmastery.com/time-series-data-stationary-python/

[9] Investopedia. (n.d.). *What Is Autocorrelation?* Investopedia. Retrieved July 6, 2021, from https://www.investopedia.com/terms/a/autocorrelation.asp

[10] Tinungki, G. M. (2019). The analysis of partial autocorrelation function in predicting maximum wind speed. *IOP Conference Series: Earth and Environmental Science*, *235*, 012097. https://doi.org/10.1088/1755-1315/235/1/012097

[11] Mitrović, D., Zeppelzauer, M., & Brietender, C. (2010, January 1). *Features for Content-Based Audio Retrieval*. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0065245810780037

[12] Habsari, H. D. P., Purnamasari, I., & Yuniarti, D. (2020). FORECASTING USES DOUBLE EXPONENTIAL SMOOTHING METHOD AND FORECASTING VERIFICATION USES TRACKING SIGNAL CONTROL CHART (CASE STUDY: IHK DATA OF EAST KALIMANTAN PROVINCE). *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *14*(1), 013–022. https://doi.org/10.30598/barekengvol14iss1pp013-022

[13] Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

[14] Glen, S. (n.d.). *RMSE: Root Mean Square Error*. Statistics How To. Retrieved July 6, 2021, from https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/

[15] Gujarati, D. N., & Porter, D. C. (2008). *Basic Econometrics* (5th ed.). McGraw-Hill Education.

[16] Choudhury, A., & Jones, J. (2014). CROP YIELD PREDICTION USING TIME SERIES MODELS. *Journal of Economic and Economic Education Research*, *15*(3), 53–68. https://www.alliedacademies.org/articles/crop-yield-prediction-using-time-series-models.pdf3