International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET2411146

Comparison of SARIMA, Bagging Exponential Smoothing with STL Decomposition and Robust STL Decomposition for Forecasting Red Chili Production

Titin Agustina[1], Anwar Fitrianto[2], Indahwati[3]

[1,2,3]Department of Statistics, IPB University, Bogor, West Java, Indonesia

Email: titinagustina@apps.ipb.ac.id

ARTICLEINFO

Article History:
Accepted : 10 March 2024
Published: 25 March 2024

Publication Issue :
Volume 11, Issue 2
March-April-2024

Page Number :
64-73

ABSTRACT

Time series analysis enables the identification of trends and patterns in data, allowing for the development of forecasting models that predict future values. One effective approach for forecasting seasonal time series data is the Seasonal Autoregressive Integrated Moving Average (SARIMA) method. Bagging Exponential Smoothing with STL Decomposition (BES-STL) is an ensemble machine learning method aimed at enhancing forecasting accuracy. STL Method, which stands for Seasonal-Trend decomposition using Loess, is utilized to decompose time series data into three components, namely trend, seasonal, and remainder components. In the remainder component, the process of bootstrap aggregation (bagging) with Moving Block Bootstrapping (MBB) is used to obtain synthetic data, followed by averaging the value by month from the entire series as the forecast results. A comparative analysis was conducted using the SARIMA, BES-STL, and BES-RSTL models. The optimal model, with the lowest MAPE and RMSE, is then implemented to predict national red chili production. The results indicate that the SARIMA(1,1,1)(0,1,1)12 model has the best performance with a MAPE of 7.06 and a RMSE of 95,473. The top-performing model is utilized to forecast data from January to December 2022. Additionally, the forecasted results are compared to the actual data, resulting in a highly precise MAPE of 5.39.

Keywords: BES-STL, MBB, Red Chili Production, SARIMA, Time Series

I. INTRODUCTION

Time series data is a collection of data points arranged by their time of observation. This data is typically collected at regular intervals. Time series analysis is a statistical method used to examine time series data and identify trends and patterns that emerge from it. One common pattern found in time series data is seasonal patterns, which are fluctuations that occur regularly within a specific time frame. Time series data analysis can be used to identify trends and patterns in the data, which can then be used to create a forecasting model based on previous observations.

One commonly used method for forecasting time series data is ARIMA (Autoregressive Integrated Moving Average), which is used to analyze time series data that has a stationary pattern in mean and variance. Time series data with seasonal patterns can be modeled using an extension of the ARIMA model called SARIMA or Seasonal ARIMA (Montgomery et al., 2015).

The accuracy of the model can be improved by using resampling techniques, such as the bootstrap aggregating (bagging) method. Bagging, invented by Breiman (1996), is a form of ensemble machine learning that involves using multiple models with the same algorithm (Mu'tashim et al., 2023).

The bagging method is a procedure in which each model is trained on a different subset of samples from the same dataset. Bergmeir et al. (2016) introduced the use of bagging to forecast time series data through resampling, known as Bagging Exponential Smoothing (BES). BES is a combination of several different Exponential Smoothing models aimed at improving forecasting accuracy. The Exponential Smoothing models utilized in BES have varying parameters, including alpha, beta, or theta, which are randomly set and tested on the same data. The final forecast is produced by combining the results of the models.

A decomposition process can provide a clearer understanding of patterns and trends in time series data. Time series data with a strong seasonal pattern can be analyzed using the STL (Seasonal-Trend decomposition using Loess) method. This method divides the time series data into three components: trend, seasonal, and remainder components. The seasonal and trend components of the data are estimated using Loess (Locally Weighted Scatterplot Smoothing) (Hyndman and Athanasopoulos, 2018). Loess is a non-parametric smoothing technique used in time series analysis to eliminate noise or random patterns in the data. This allows for clearer observation of seasonal patterns and trends in the data. The process of decomposition proceeds by analyzing these components individually and then reintegrating them to conduct forecasting.

RSTL (RobustSTL) is a time series decomposition technique that extends the STL method (Wen et al. 2019). It considers the presence of outliers or extreme data in time series, resulting in more accurate forecasts than STL. RSTL is suitable for processing time series data with seasonal patterns and outliers, such as factory production or retail sales data.

Agricultural production data for certain commodities, such as chili, are subject to seasonal patterns, fluctuations, and outliers. The production of fresh chili plants is influenced by weather conditions, with plants in the rainy season being susceptible to pests and diseases that can cause rotting and failure to harvest. This, in turn, impacts the market price of chili due to reduced supply. The price of chili tends to fall when there is a bumper harvest or at the end of the dry season due to abundant supply, as noted by Kementan (2020). Outliers may occur if the government intervenes to increase chili production to meet the community's needs. In Indonesia, chili is considered one of the most important agricultural products due to its high public demand, making it a strategic commodity that requires special attention from the government. It is important to note that this information is objective and free from any biased or emotional language. Also, in Indonesia chili is considered one of the most important agricultural products due to its high public demand, making it a strategic commodity that requires special attention from the government. The Ministry of Agriculture's Strategic Plan for 2020-2024 includes chili as one of the seven main food commodities, alongside rice, corn, soybeans, onions, sugar cane, and beef/buffalo. According to BPS-Statistics Indonesia data from 2020, national red chili production has shown an increasing trend from 2015 to 2019, with a growth rate of 4.02% per year. During the same period, the producer-level

price of red chili increased significantly, at a rate of 10.57% per year. Ensuring an adequate supply of red chili is crucial for maintaining price stability.

The data on national red chili production is presented in a time series format. The use of time series analysis can aid in predicting future national red chili production. The selection of a forecasting model when analyzing time series data can impact the accuracy and validity of the forecast results and the decisions made based on them. This paper compares the performance of SARIMA, BES-STL, and BES-RSTL to determine the best model for forecasting.

II. METHODS AND MATERIAL

The study utilized secondary data, specifically red chili production in quintals obtained from Statistics Indonesia and the Ministry of Agriculture. The data consists of monthly national-level records from January 2013 to December 2021, divided into training and testing data. The candidate model is created using training data from January 2013 to December 2019, while testing data is used to validate the best model from January 2020 to December 2021. Additional data from January to December 2022 is used to compare real and forecasted data and evaluate the accuracy of the forecast.

SARIMA

The SARIMA (Seasonal Autoregressive Integrated Moving Average) method is utilized to analyze data with a seasonal pattern, which repeats over a fixed interval of time. The SARIMA(p,d,q)(P,D,Q)s model is formed by determining the appropriate order for each model. The SARIMA model, as denoted by Montgomery et al. (2015), is as follows:

$$\Phi_P(B^S)\phi_p(B)(1-B)^d(1-B^S)^D Y_t = \delta + \theta_q(B)\Theta_Q(B^S)\varepsilon_t \quad (1)$$

where:

p, d, dan q	: non seasonal orders
P, D, dan Q	: seasonal orders
Y_t	: time series data at t period
δ	: constant
ε_t	: lag in t period
B	: backshift operator
S	: seasonal period
Φ_P	: seasonal AR component
Θ_Q	: seasonal MA component

The necessary steps are:

1. Check the stationarity of the data.
2. Divide the data into training and testing sets.
3. Identify the model by analyzing its non-seasonal and seasonal orders using ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots, as well as its differentiation value.
4. Determine the provisional estimation parameters based on the predetermined order.
5. Test the significance of the parameters.
6. Perform a model diagnostic test on the residuals.
7. Model overfitting.
8. Choose the best model.
9. Validate the model with test data.

BES-STL

BES is a combination of several different Exponential Smoothing models that aim to improve forecasting accuracy. The results of these models are combined to produce the final forecast. Follow these steps:

1. Box-Cox Transformation

This represents a commonly used conversion to stabilize the variability of a time series, initially suggested by Box and Cox (1964). The definition is as follows:

$$y_t(\lambda) = \begin{cases} ln\, y_t & \lambda = 0 \\ (y_t^{\lambda} - 1)/\lambda, & \lambda = 0 \end{cases} \quad (2)$$

In this equation, y_t represents the t-th time series and λ is the transformation parameter, which has a value between 0 and 1.

2. STL Decomposition

STL decomposition (Seasonal-Trend decomposition using Loess) method is employed to divide time series data into three components: trend, seasonal, and remainder components.

3. Moving Block Bootstrapping the remainder

The remainder components are generated 100 times using Moving Block Bootstrapping (MBB) with block sizes of $l = 24$. Afterward, the trend and seasonal components that have been broken down are added back with 100 new remainder components. The resulting series are then back-transformed to obtain 100 new series.

4. Holt's Winter

The Holt Winters forecasting method was applied to each series to generate 100 forecasts. It is important to note that there are two Holt Winter methods available, the multiplicative and additive methods.

The Multiplicative Holt Winters model is obtained through:

$$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + T_{t-1}) \quad (3)$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \quad (4)$$

$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma)S_{t-s} \quad (5)$$

$$F_{t+m} = (L_t + T_t m)S_{t-s+m} \quad (6)$$

The Holt Winters additive model has the following equation:

$$L_t = \alpha(Y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + bT_{t-1}) \quad (7)$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \quad (8)$$

$$S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-s} \quad (9)$$

$$F_{t+m} = L_t + T_t m + S_{t-s+m} \quad (10)$$

where L_t, T_t, and S_t are the level, trend and seasonal components at time *t, s* is the seasonal period, and *m* is the forecasting period. While α, β, and γ is the smoothing parameter.

5. Aggregation

The average value of the entire forecast series in each month is taken as the final value of the forecast.

BES-RSTL

BES-RSTL is basically the same as BES-STL, the difference being the consideration of outliers in the analysis. In this method, the decomposition process is performed with robustness in mind, that is, by minimizing the influence of outlier data in the decomposition process. The algorithm used for the decomposition in RSTL can be found in Wen et al. (2019) in a journal titled "RobustSTL: A Robust Seasonal-Trend Decomposition Algorithm for Long Time Series".

Selection of Best Model

The best model selection among SARIMA, BES-STL and BES-RSTL is determined based on the lowest values of Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). The MAPE is given by the formula:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100\%$$

While RMSE with the equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

where $\hat{y}_t$ is the predicted value at time *t*, y_t is the actual value at time *t*, and *n* is the amount of data.

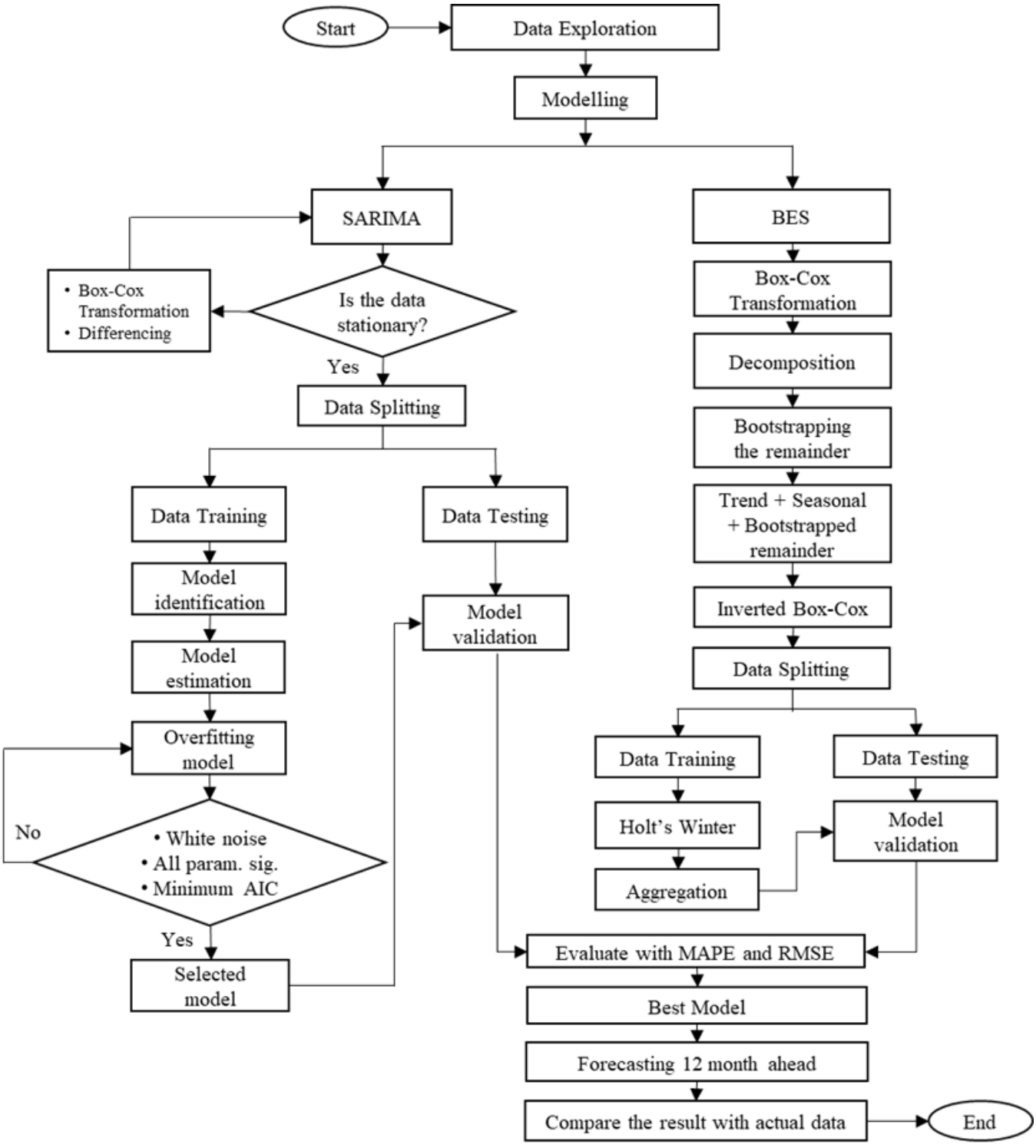

Figure 1. SARIMA and BES flowchart

The forecasting results with the best model are then compared to the actual data by looking at the MAPE value. The interpretation of the MAPE is as follows:

Table 1
Interpretation of MAPE value

MAPE	Interpretation
<10%	Highly accurate

10-20%	Accurate
20-50%	Less accurate
>50%	Inaccurate

III.RESULTS AND DISCUSSION

Exploratory Data Analysis

The first stage of data analysis was conducted by examining national red chili production data from January 2013 to December 2021. The data used to build the model are red chili production data from January 2013 to December 2019, while data from January 2020 to December 2021 are used for validation.

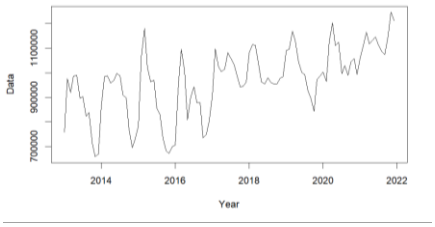

Figure 2. Time series plot of red chili production

There are 108 data points used in the modeling, representing the amount of national red chili production for eight years (Fig. 2), where the X-axis shows the observation time, while the Y-axis shows the amount of red chili production in quintals.

The results from Fig. 2 show that the national red chili production data has a seasonal pattern with a 12-month period. The plot of the national red chili production data during 2013-2021 shows significant fluctuations, with an almost similar trend each year. Red chili production tends to be low in January, but increases dramatically in February, which is the beginning of the main red chili harvest. Furthermore, red chili production shows a downward trend until it reaches its lowest point in October and November, except in 2021, when it reaches its highest point in November.

National red chili production data from January 2013 to December 2021 has the highest production in November 2021. Conversely, the lowest national red chili production occurred in November 2013. This decline in production may be caused by weather factors such as the rainy season and natural phenomena in the form of the eruption of Mount Sinabung in the province of North Sumatra, which is one of Indonesia's red chili center provinces. According to BMKG forecasts, October and November 2013 were the beginning of the rainy season for a number of regions in Indonesia. Climate change, such as rainfall, can cause a decrease in red chili production.

Further exploration is done by creating a line box or box plot per month to see if the data has outliers or not. In Fig. 3, there are 3 small circle symbols showing that there are 3 outlier points, namely in month 5 and month 7. If we follow further, these outliers occurred in May 2016, July 2015 and July 2017. In this case, the outlier data is not removed, or in other words, it is still included in the analysis.

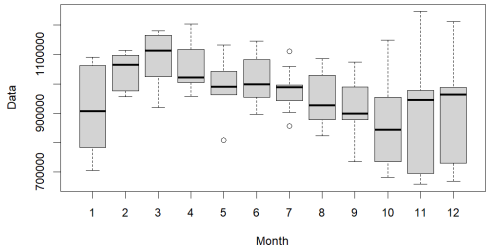

Figure 3. Box plot by month

SARIMA Model

To build a SARIMA model, it is necessary to test the stationarity of the overall data. Based on the time series plot in Fig. 4, the data may not be stationary in either mean or variance.

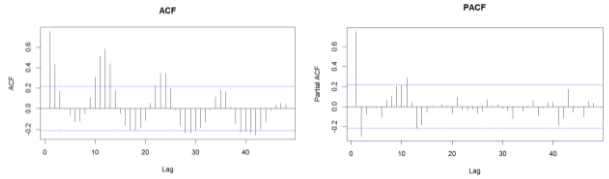

Figure 4. ACF and PACF Plot of training data

To be sure, a formal test of the assumptions was performed. The Augmented Dicky Fuller (ADF) test is used to test the stationarity of the data in the mean. Data is said to be stationary in the mean if the ADF test produces a p-value smaller than 0.05.

Table 2
ADF Test Results

Component	P-value	Description
Non Seasonal	0.01	Stationary
Seasonal	0.6962	Not stationary

Table 2 shows that red chili production has reached stationary conditions for non-seasonal components, but has not yet reached stationary conditions for seasonal components. Furthermore, the stationarity of the variance is checked with the Box-Cox transformation. The λ value obtained is 0.9999 or close to 1, so it can be concluded that the data are stationary in variety and no transformation is performed.

The overall data is then divided into training data and test data. Model identification is done on the training data using ACF and PACF plots. Fig. 5 shows that with differencing 1, the ACF value is cut off at lag 2, indicating that the order of the Moving Average (MA(q)) model is 2. The PACF value is also cut off at lag 2, indicating that the order of the Auto Regressive (AR(p)) model is 2. The candidate or preliminary models are $(2,1,0)(0,1,2)^{12}$ and $(0,1,2)(2,1,0)^{12}$.

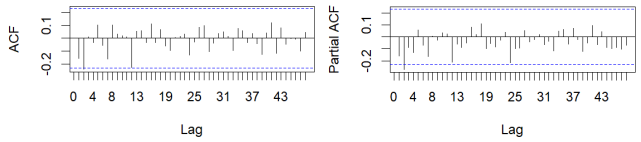

Figure 5. ACF and PACF plot of stationary training data

The obtained preliminary model is then tested for parameter significance and white noise for freedom and normality of residuals. Model parameters are considered significant if they have a value less than α (0.05). The Ljung-Box and Jarque Bera tests are used to test for freedom and normality, where the model is said to be white noise if it has a p-value greater than α (0.05). Model overfitting is done by adding orders to the existing tentative models. When the models have met the assumptions of all significant parameters and white noise, the AIC value is calculated. The best model is the one that satisfies the assumptions of white noise, all parameters are significant, and has a minimum AIC. The results of the preliminary model test and the overfitting are below:

Table 3
Test Results of the Preliminary and the Overfitting Model

Model SARIMA	**Ljung-Box****	**Jarque Bera**	**AIC**
Preliminary			
$(2,1,0)(0,1,2)^{12}$	Sig.	0.3147	1792,89
$(0,1,2)(2,1,0)^{12*}$	Sig.	0.1795	1792,04
Overfitting			
$(0,1,2)(0,1,2)^{12}$	Sig.	0.2859	1790,89
$(2,1,2)(0,1,2)^{12}$	Sig.	0.4425	1792,24
$(1,1,1)(0,1,2)^{12}$	Sig.	0.5355	1789,03
$(1,1,1)(0,1,1)^{12*}$	**Sig.**	**0.1657**	**1790,49**

* : All parameters significant at α=0.05

**: At lag 5, 10, 15, 20, 25, 30

Based on the best SARIMA model, a production forecast is made for the next 12 periods. Fig. 6 shows a forecast plot whose pattern resembles the actual data. The forecast shows the highest forecast in March 2021 of 1,216,0006 quintals or 121.6 thousand tons and the lowest forecast in October 2020 of 928,435.5 quintals or 92.84 thousand tons. Fig. 6 also shows that the forecast pattern tends to follow the actual production pattern well.

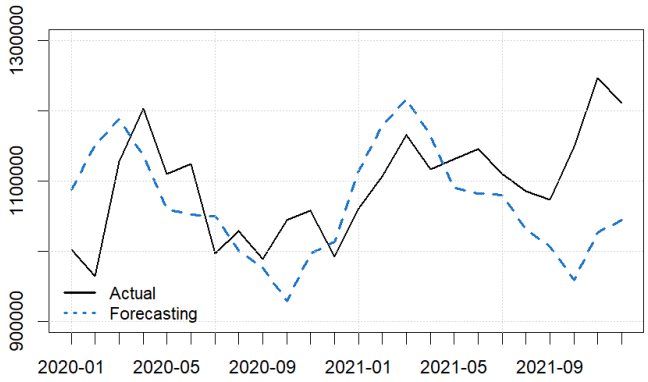

Figure 6. Plot of actual and forecast data 2020 – 2021 with SARIMA

BES-STL Model

There are several steps to performing BES-STL on time series data. First, a Box-Cox transformation is performed to stabilize the variance. The next step is to decompose the data using the STL method because the time series data has an indication of seasonality. This step decomposes the data into trend, seasonal and remainder components, which can be seen in Fig. 7.

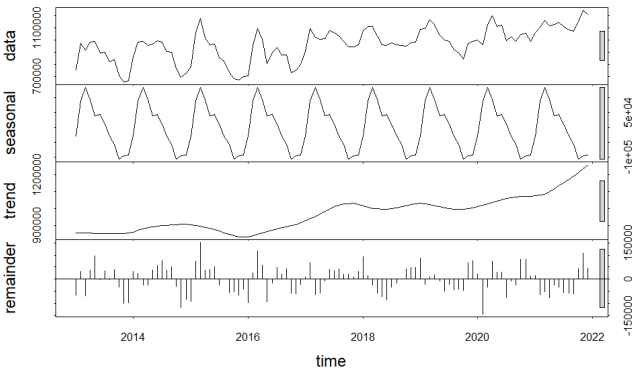

Figure 7. Decomposition Plot

After decomposition, the remainder components are used for the bootstrapping process. To create 100 new data series, 99 data are generated from the original data. The similarity between the bootstrapped series and the original series depends on the behavior of the remaining components. If the residual component has a small variance, this process will produce synthetic series that are similar to each other. The results of the bootstrapping method with MBB are shown in Fig. 8, where the original series are identified by black, while the synthetic series are identified by various colors other than black.

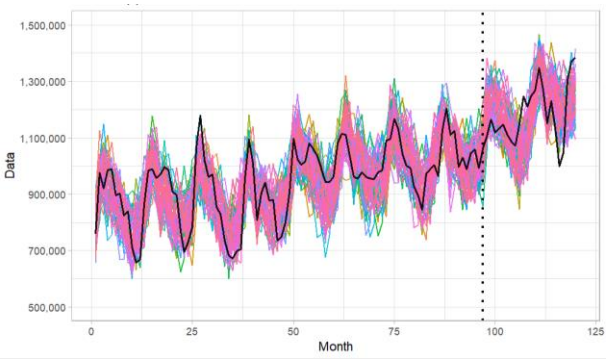

Figure 8. Bootstrapped versions (various colors) of the original series (black)

Furthermore, Holt's Winter method is applied to predict each bootstrapped training data series, and then the average of each month is taken. The visualization between the forecasting data and the test data is shown in the following figure.

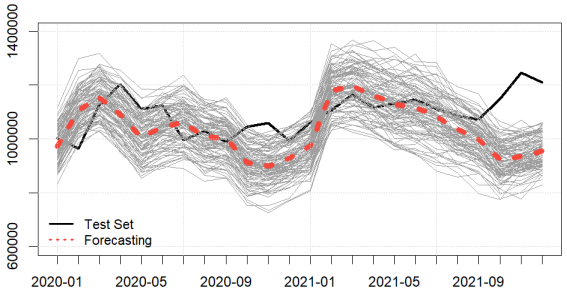

Figure 9. Forecasting and Test Data with BES-STL

BES-RSTL Model

As in the case of BES-STL, a decomposition is first performed, but this time taking into account the outliers. The resulting remainder is then bootstrapped with MBB to generate 99 new series. The average is used as the final forecast result. Below is an overview of the forecast results using the test data.

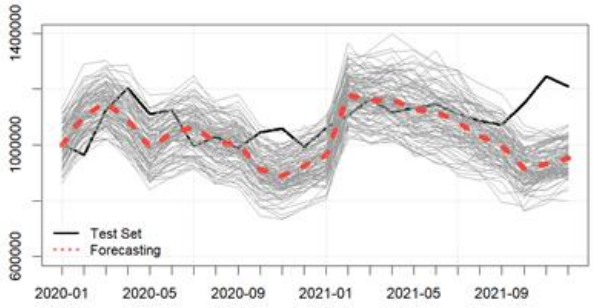

Figure 10. Plot of Forecasting and Test Data with BES-RSTL

Model Evaluation

The forecast accuracy results obtained from the test data are used to compare the performance of the models. A comparison of MAPE and RMSE values is used to determine which method provides the best forecast results among the methods.

Table 4
Model Accuracy

Method	MAPE	RMSE
SARIMA(1,1,1)(0,1,1)^{12}	**7,06**	**95.473,93**
BES-STL	8,09	119.673,93
BES-RSTL	8,08	121.766,22

Considering several models that have been made, it can be concluded that in the national red chili production data, the SARIMA(1,1,1)(0,1,1)^{12} model has a better performance. This is because the model has lower MAPE and RMSE values than other models.

Comparison between Forecasting Results and Actual Values

The model with the best performance is then used to forecast data for the period January to December 2022. The forecast results were then compared with the actual data. The empirical results summarized in Table 5 and an average MAPE of 5.39 was obtained.

Table 5
MAPE of forecasting results with actual data

Time	MAPE
Jan-22	2.17
Feb-22	2.74
Mar-22	2.72
Apr-22	1.82
May-22	6.03
Jun-22	5.36
Jul-22	3.25
Aug-22	11.56
Sep-22	3.36
Oct-22	15.84
Nov-22	3.33
Dec-22	6.56

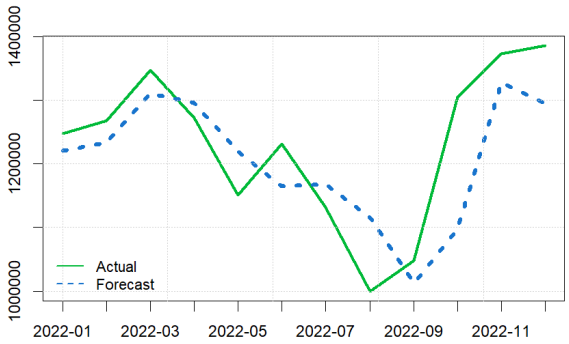

Figure 11. Visualizing actual and forecast data in 2022 using SARIMA(1,1,1)(0,1,1)^{12} method

IV.CONCLUSION

The results of the comparison of the SARIMA model, Bagging Exponential Smoothing with STL Decomposition and Robust STL Decomposition show that SARIMA has better accuracy compared to other models on national red chili production data. Therefore, the forecast for the period from January to December 2022 is made using the SARIMA(1,1,1)(0,1,1)^{12} model. The comparison of the forecast results with the actual data showed a MAPE of 5.39 which can be said to have a very high degree of accuracy.

V. REFERENCES

[1] Bergmeir C, Hyndman RJ, Benitez JM. 2016. Bagging exponential smoothing methods using STL decomposition and Box-Cox transformation. International Journal of Forecasting. 32(2016):303-312. doi:10.1016/j.ijforecast.2015.07.002.

[2] BPS. 2020. Statistik Indonesia, Statistical Yearbook of Indonesia 2020. Jakarta: BPS.

[3] Hyndman RJ, Athanasopoulos G. 2018. Forecasting: Principles and Practice. 2nd ed. Australia: OTexts.

[4] Kementan. 2020. Outlook Cabai 2020. Jakarta: Kementan.

[5] Maulana HA. 2018. Pemodelan deret waktu dan peramalan curah hujan pada dua belas stasiun di Bogor. Jurnal Matematika, Statistika dan Komputasi. 15(1):50-63. doi: 10.20956/jmsk.v15i1.4424.

[6] Montgomery DC, Jennings CL, Kulahci M. 2015. Introduction to Time Series Analysis and Forecasting. 2nd ed. New Jersey: John Wiley & Sons, Inc.

[7] Mu'tashim ML, Zaidiah A, Yulistiawan BS. 2023. Klasifikasi ketepatan lama studi mahasiswa dengan algoritme random forest dan gradient boosting (Studi kasus fakultas ilmu komputer Universitas Pembangunan Nasional Veteran Jakarta). In Proceedings of the Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya (Senamika). 4(1):155-166.

[8] Wen Q, Gao J, Song X, Sun L, Xu H, Zhu S. 2019. RobustSTL: A robust seasonal-trend decomposition algorithm for long time series. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19). 33(01): 5409-5416.