# Data Prediction For Coffee Harvest Using Least Square Method

# Data Prediction For Coffee Harvest Using Least Square Method

**Edi Surya Negara, Keni Keni*, Ria Andryani**
Informatics Department, Faculty of Computer Science, Universitas Bina Darma
Faculty economics & Business, Universitas Tarumanagara
Information System Department, Faculty of Computer Science, Universitas Bina Darma

e.s.negara@binadarma.ac.id, *keni@fe.untar.ac.id, ria.andryani@binadarma.ac.id

**Abstract.** Pagaralam is one of the highest quality coffee producing regions in Indonesia. But the problem that is often found by farmers is the lack of knowledge and predictions about the coffee harvest they will produce in the next period. The solution that can be given is developing an application to be able to analyze and predict coffee yield data for the next harvest period. This study produces a calculation using the Least Square method which can produce a prediction algorithm for coffee yields with the lowest prediction error rate with an MPE of 13.72 and the greatest accuracy using a MAPE of 0.0166 which is implemented in a Coffee Harvest Prediction Application.

## 1. Introduction

Indonesia as one of the best and biggest coffee-producing countries in the world should have increased its productivity. Especially now that Indonesia has begun implementing the 4.0 industrial revolution by designing a roadmap namely Making Indonesia 4.0. Information technology can make it easier for humans to do their jobs. Thus, information technology can be implemented in various fields, one of which is through the use of data analytics and predictions such as the use of big data in agriculture [1][2][3] and the development of agricultural applications using data mining [4] [5]. The linkage to the coffee commodity industry in Kota Pagaralam is that it is time to increase productivity in implementing the 4.0 industrial revolution with efforts to facilitate the process between farmers and consumers in interacting by utilizing digitization as a means and obtaining predictions of coffee yields for the next period. But the problem that often occurs is the availability of coffee in certain periods is sometimes not in accordance with the high demand, and vice versa which makes the level of coffee production in Indonesia sometimes become unstable. Accurate information is needed in daily life, information will become an important element in the development of society today and in the future. Utilizing existing data in information systems to support decision-making activities is not enough just to rely on operational data, but we need data analysis to explore the potential of existing information [4]. Therefore we need an application as prediction information on coffee yields with the hope that the yield in the previous period can be a benchmark in predicting the yield in the next period.

There are many methods that can be used to predict coffee yields that have been developed [5]. But this article presents a complete analysis and application design using the Least Square Method in a case study of coffee commodity results in Kota Pagaralam. The use of the Least Square method is considered good and suitable for predicting coffee yields, this algorithm is also one part of Time Series forecasting which is used to see trends from a time series data using previous time data [6][7]. In this case, it will be more focused on discussing time series analysis with the Least Square method which is divided into two cases namely odd and even data and the design of prediction applications of coffee yields.

## 2. Literature

### *2.1. Algorithm Prediction*

Prediction is what will happen in the future, while the plan, is the determination of what will be done in the future. The predicted problems vary such as rainfall estimates, possible winners in the elections, match scores, or inflation rates. Based on this understanding, prediction can be an attempt or activity

to predict an activity that will occur in the future by using a test from the past. Prediction methods there are several types according to their nature namely [8]:

1. Qualitative (Non-Statistical Method or Opinion Method) Is a way of forecasting or forecasting based on past data, current data, targets to be achieved, based on one's intuition and experience and opinions.
2. Quantitative (Statistical Method) Is a forecasting method that focuses on numerical calculations using a variety of statistical methods. By using quantitative forecasting methods such as this it is expected to eliminate subjective elements or personal opinions as far as possible, so that the estimated results are more accurate and can be accounted for, and can be more useful as a basis for company decision making.

There are several methods commonly used in time series analysis. These are as follows: Free Hand Method, Semi Average Method, Moving Average Method, Least Square Method, Moment Method, Parabolic Method [8].

### *2.2. Least Square Method*

Least Square method is a forecasting method used to see trends from time series data [9]. Least Square Method: The method used for time series analysis is the Free Linear Line Method (Free Hand Method), the Semi Average Method, the Moving Average Method, and the Least Squared Method Square Method) [10][11]. The formulas used are:

$$\boldsymbol{Y = a + bx} \qquad \textbf{(1)}$$

*Y* = Total Prediction / Calculation
a and b = Coefficient
x/t = A certain time in the form of a code

Furthermore, to find out the coefficients a and b can searched using the formula:

$$\boldsymbol{a = \frac{\sum Y}{n}} \qquad \textbf{(2)}$$

$$\boldsymbol{b = \frac{\sum tY}{\sum t^2}} \qquad \textbf{(3)}$$

To make the value of t2 = 0 the sum of the year data being measured is even and odd [7]. With guidance as follows :

1. If the amount of year data does not expire divided by two namely odd then used scale x = 1 year. As well as the year on which to base placed in the middle of the year.
2. If the amount of data spent in years is divided by two, even used scale x = 1/2 year. As well as the year on which to base is the middle year.

### *2.3. Prediction Testing Method*

The calculation of the results of predictions made in each case will not likely be absolutely correct, in accordance with the actual realization that will occur [12]. Often the difference occurs in the results obtained, this is usually called an error or error. through this error, the value can be done several types of analysis so that it can compare which forecasting method best suits the data held and how well the method used is. This can be seen from the comparison between the error values produced by each method. To find out the value of errors in forecasting can use some statistical testing. This indicator aims to choose the method that has the smallest deviation [13][14]. The following prediction testing methods will be used:

MAPE (Mean Absolute Percentage Error)

$$\boldsymbol{MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Yt-\hat{Y}t|}{Yt}\, x\, 100\%} \qquad \textbf{(4)}$$

MPE (Mean Percentage Error)

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{Yt-\hat{Y}t}{Yt}\ x\ 100\% \qquad (5)$$

## 3. Research Methodology

The flow of this research refers to the research framework as in Figure 1, namely: Data collection, Application of algorithms, Evaluation of calculation results, Application design, Feasibility test, Application evaluation, and Experimental documentation.

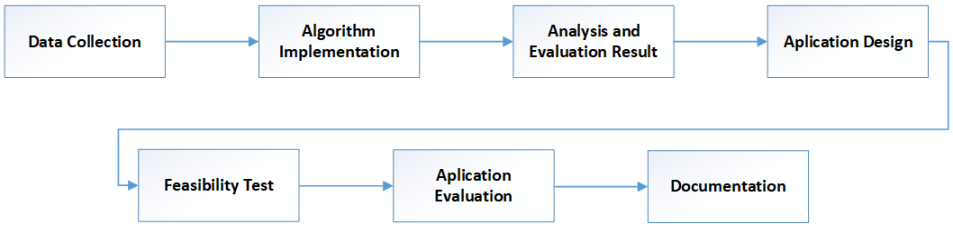


Figure 1. Research Stages

### *3.1. Method of collecting data*

The study used statistical data on coffee yields from Pagaralam City by comparison of each district or village in 2016-2018. The data contains coffee yields per sub-district or village in Pagaralam City with the division of harvest time per quarter which is 4 years or 4 quarters for 1 year. Coffee yield data generated is using tonnage units. The data collection methods used are as follows:

**Observation.** In this observation method, the researcher makes observations and data retrieval directly by visiting the Department of Agriculture of Pagaralam City to get the data needed as a first step to analyzing predictive objects, the data observed and analyzed also constitutes coffee yield data from each village or village. In Pagaralam City, the researchers did not stop there, the observation continued by visiting several coffee fields in two different districts as samples to see firsthand the similarity of conditions with data obtained from the reality in the field.

**Interview.** The interview method was conducted by researchers while in the Pagaralam City Agriculture Office, interviews were conducted to find information about the distribution of the existing harvest period in the City of Pagaralam itself in this case referring to coffee commodities as well as questions about the required coffee yield data. After that, interviews were also conducted with coffee farmers, researchers conducted in 2 coffee fields in 2 different districts, in this interview the researchers asked about the harvest time they did and how much yield they get in each period.

**Literature study.** In this literature study researchers look for and read books, journals, or articles related to predictions of coffee yields, as well as understand and explore various scientific theories related to the object of research conducted.

## 4. Results and discussion

The following is coffee yield data for Pagaralam City, Penjakung Urban Village, I Quarter 2016 - Quarter IV 2018 in table 1. In this study, the prediction of coffee yields in the first quarter of 2019 will be calculated using Least Square. For the selection of even and odd data types. Then the accuracy of prediction needs to be tested, in this case using the calculation of MPE and MAPE.

Table 1. Data on Harvest Coffee Results for Quarter I 2016 - Quarter IV 2018

| Year | Time | Harvest Data |
|---|---|---|
| 2016 | Triwulan I | 500 |
| 2016 | Triwulan II | 190 |
| 2016 | Triwulan III | 237 |
| 2016 | Triwulan IV | 260 |
| 2017 | Triwulan I | 206 |

| | | |
|---|---|---|
| 2017 | Triwulan II | 75 |
| 2017 | Triwulan III | 94 |
| 2017 | Triwulan IV | 95 |
| 2018 | Triwulan I | 500 |
| 2018 | Triwulan II | 610 |
| 2018 | Triwulan III | 552 |
| 2018 | Triwulan IV | 260 |

***Predictions using even data***

Prediction of the yield of Penjalang coffee in the first quarter of 2019, using yield data from the first quarter of 2016 to the fourth quarter of 2018 with an even data type of 12 months.

Table 2. Calculation of the Even Least Square Data Formula

| N | Year | Time | Harvest (Y) | Coding (t) | tY | $t^2$ |
|---|---|---|---|---|---|---|
| 1 | 2016 | Triwulan I | 500 | -11 | -5500 | 121 |
| 2 | 2016 | Triwulan II | 190 | -9 | -1710 | 81 |
| 3 | 2016 | Triwulan III | 237 | -7 | -1659 | 49 |
| 4 | 2016 | Triwulan IV | 260 | -5 | -1300 | 25 |
| 5 | 2017 | Triwulan I | 206 | -3 | -618 | 9 |
| 6 | 2017 | Triwulan II | 75 | -1 | -75 | 1 |
| 7 | 2017 | Triwulan III | 94 | 1 | 94 | 1 |
| 8 | 2017 | Triwulan IV | 95 | 3 | 285 | 9 |
| 9 | 2018 | Triwulan I | 500 | 5 | 2500 | 25 |
| 10 | 2018 | Triwulan II | 610 | 7 | 4270 | 49 |
| 11 | 2018 | Triwulan III | 552 | 9 | 4968 | 81 |
| 12 | 2018 | Triwulan IV | 260 | 11 | 2860 | 121 |
| **Jumlah** | | | **$\sum Y = 3579$** | - | **$\sum tY = 4115$** | **$\sum t^2 = 572$** |

Where *Y* Yield Data, *a* and *b* Timeline Data, $\sum Y$ Amount of Harvest Data, $\sum t$ Amount of scoring or time period, $\sum tY$ Amount of Harvest Data *x* Amount of time period, *n* Number of time period data. The first step to find the values of *a* and *b* uses the formula below:

$$a = \frac{\sum Y}{n} = \frac{3579}{12} = 298{,}25$$

$$b = \frac{\sum tY}{\sum t^2} = \frac{4115}{572} = 7{,}1940$$

The next step is to find the predicted *Y* value using the following formula:

$Y = a + bx$

$Y = 391{,}772$

So the predicted yield of the first quarter coffee yield of 2019 is equal to = **391,7727 tons.**

To find out the accuracy of predictions, it is necessary to test the smallest error rate. Testing the error using the prediction accuracy test with MPE and MAPE.

MAPE = 0,0166 so the error rate (error) in the prediction 0,0166

MPE = 13,72 so the error rate in the prediction13,72

***Predictions using odd data***

Prediction of the yield of Penjalang coffee in the first quarter of 2019, using yield data from the first quarter of 2016 to the third quarter of 2018 with an even data type of 11 months.

Table 3 Calculation of the Odd Least Square Data Formulas

| N | Year | Time | Harvest (Y) | Coding (t) | tY | $t^2$ |
|---|---|---|---|---|---|---|
| 1 | 2016 | Triwulan I | 500 | -5 | -2500 | 25 |
| 2 | 2016 | Triwulan II | 190 | -4 | -760 | 16 |
| 3 | 2016 | Triwulan III | 237 | -3 | -711 | 9 |
| 4 | 2016 | Triwulan IV | 260 | -2 | -520 | 4 |
| 5 | 2017 | Triwulan I | 206 | -1 | -206 | 1 |
| 6 | 2017 | Triwulan II | 75 | 0 | 0 | 0 |
| 7 | 2017 | Triwulan III | 94 | 1 | 94 | 1 |
| 8 | 2017 | Triwulan IV | 95 | 2 | 190 | 4 |
| 9 | 2018 | Triwulan I | 500 | 3 | 1500 | 9 |
| 10 | 2018 | Triwulan II | 610 | 4 | 2440 | 16 |
| 11 | 2018 | Triwulan III | 552 | 5 | 2760 | 25 |
| **Jumlah** | | | **$\sum Y = 3319$** | **-** | **$\sum tY = 2287$** | **$\sum t^2 = 110$** |

Where *Y* Yield Data, *a* and *b* Timeline Data, $\sum Y$ Amount of Harvest Data, $\sum t$ Amount of scoring or time period, $\sum tY$ Amount of Harvest Data *x* Amount of time period, *n* Number of time period data
The first step to find the values of a and b uses the formula below:

$$a = \frac{\sum Y}{n} = \frac{3319}{11} = 301{,}7273$$
$$b = \frac{\sum tY}{\sum t^2} = \frac{2287}{110} = 20{,}791$$

The next step is to find the predicted *Y* value using the following formula:
$Y = a + bx$
$Y = 572{,}00$

So the predicted yield of the first quarter coffee yield of 2019 will be equal to = **572.00 tons.**
To find out the accuracy of predictions, it is necessary to test the smallest error rate. Testing the error using the prediction accuracy test with MPE and MAPE.

MAPE = -0,0110 so the error rate in the prediction -0,0110
MPE = 16,4923 so the error rate (error) in the prediction 16,4923

From the prediction results of yields in the first Quarter 2019 period above using Even and Odd Data, the accuracy of forecasting results is shown in Even data types with the smallest error rate ie MPE of 13.72 and the greatest accuracy using MAPE of 0.0166. Then it can be suggested for forecasting or prediction of the yield of the coffee of the crossing in the first quarter of 2019 it would be better to use yield data with even data types.

### *4.1. Application Design*

The prediction application of Pagaralam coffee yield as seen in the use case application in Figure 2 involves 2 actors consisting of user and admin. The two actors have their respective functions and duties on the system, the User is a general user who needs information related to coffee commodities in Pagar Alam City, while the Admin is a farmer or policymaker who can manage the system as a whole. Therefore, the application is made login only for the admin to be able to manage the

information on the system, while the user does not need to log in when entering the system, but can only see not to manage. Then for the application class diagram can be seen in Figure 2 below.

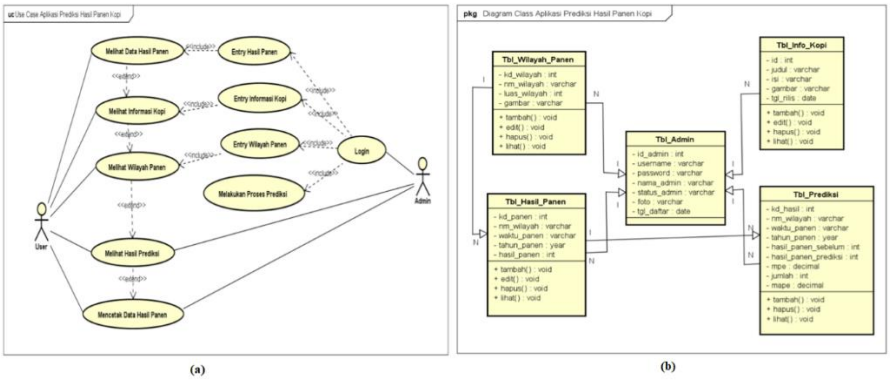


Figure 2. (a) Use Case Application; (b) Application Class Diagram

### *4.2. Prediction Error Test*

After conducting the analysis phase using the calculation algorithm, the Least Square method discussed earlier, it can be stated that the prediction object of this coffee harvest can be carried out and can be continued with implementation in an application where the design has been produced in a real form and situation ie Application Prediction of City Harvest Coffee Products. With this application, it is expected to be able to help both policymakers related to the commodity coffee city of Pagar Alam or farmers, in this case, to predict their coffee yields in the next period, so that later it can help increase their market share, and can also be information for the public and policymakers to determine the planned stock inventory needs in accordance with predictions of future demand.

Table 4 Prediction Error Test Results

| Nama Wilayah | Hasil Prediksi | Uji Error | |
|---|---|---|---|
| | | MPE | MAPE |
| Penjalang | 391,7727 Ton | 13,72 % | 0,0166 |

In this research, it is said that it succeeded in making a prediction of coffee yield using coffee yield variables in the previous period by using the least square method and by testing MPE and MAPE which have accuracy levels below 10 and 20% and then have a difference with some samples of the true value of For only 10 - 10 tons, see Table 4 and Figure 3. This research can be continued by adding other variables in its determination to be more accurate in predicting the results, and certainly by using prediction algorithms that support many variables.

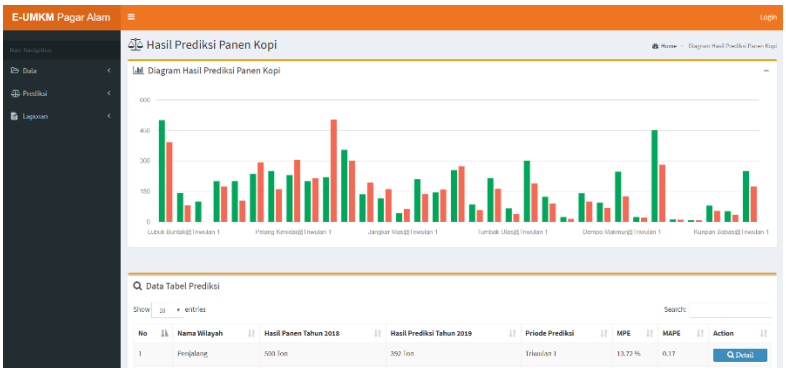


Figure 3. Regional Prediction Results

## 5. Conclusion

The use of the calculation method using the least square algorithm can be implemented in the prediction of coffee yields in the City of Pagar Alam with the lowest prediction error rate with MPE of 13.72 and the greatest accuracy using MAPE of 0.0166 which is implemented in a Coffee Harvest Prediction Application. The prediction is done using coffee yield data in the previous period, but there are some obstacles that make this study less than perfect, namely the coffee yield data of the City of Pagaralam obtained and successfully analyzed only in the last 3 years because the more data analyzed the more accurate the results will be achieved. Then theoretically the least square algorithm does not allow the addition of variables other than the yield to be analyzed to make more accurate the desired results. The suggestion to continue this research is, you should look for more data such as the scale of the past 5-10 years, then use a more conventional algorithmic method and can be analyzed using many variables.

## 6. References

[1] Kamilaris, Andreas, Andreas Kartakoullis, and Francesc X. Prenafeta-Boldú. "A review on the practice of big data analysis in agriculture." Computers and Electronics in Agriculture 143 (2017): 23-37.
[2] Pham, X. and Stack, M., 2018. How data analytics is transforming agriculture. Business Horizons, 61(1), pp.125-133.
[3] Elijah, O., Rahman, T.A., Orikumhi, I., Leow, C.Y. and Hindia, M.N., 2018. An overview of Internet of Things (IoT) and data analytics in agriculture: Benefits and challenges. IEEE Internet of Things Journal, 5(5), pp.3758-3773.
[4] Negara, E.S., Andryani, R. and Saksono, P.H., 2016. Analisis Data Twitter: Ekstraksi dan Analisis Data Geospasial. INKOM Journal, 10(1), pp.27-36.
[5] Kaur, Manpreet, Heena Gulati, and Harish Kundra. "Data mining in Agriculture on crop price prediction: Techniques and Applications." International Journal of Computer Applications 99, no. 12 (2014): 1-3.
[6] Nagini, S., Kanth, T.R. and Kiranmayee, B.V., 2016, December. Agriculture yield prediction using predictive analytic techniques. In 2016 2nd International Conference on Contemporary Computing and Informatics (ic3i) (pp. 783-788). IEEE.
[7] Cunningham, Sally Jo, and Geoffrey Holmes. "Developing innovative applications in agriculture using data mining." In The proceedings of the Southeast Asia regional computer confederation conference, pp. (1999): 25-29
[8] Guo, Hui, H. P. Liu, and Ling Wang. "Method for selecting parameters of least squares support vector machines and application." Journal of System Simulation 18, no. 7 (2006): 2033-2036.
[9] Skretting, Karl, and Kjersti Engan. "Recursive least squares dictionary learning algorithm." IEEE Transactions on Signal Processing 58, no. 4 (2010): 2121-2130.
[10] Björck, Å., 1996. Numerical methods for least squares problems. Society for Industrial and Applied Mathematics.
[11] Engel, Y., Mannor, S. and Meir, R., 2004. The kernel recursive least-squares algorithm. IEEE Transactions on signal processing, 52(8), pp.2275-2285.
[12] Jianshan, L., Changming, W., Aijun, Z. and Xiaomin, X., 2012. Residual GM (1, 1) Model-Based Prediction Method for Chaotic Time Series. Journal of Grey System, 24(4).
[13] De Myttenaere, Arnaud, Boris Golden, Bénédicte Le Grand, and Fabrice Rossi. "Mean absolute percentage error for regression models." Neurocomputing 192 (2016): 38-48.
[14] Niedbała, G., 2019. Simple model based on artificial neural network for early prediction and simulation winter rapeseed yield. Journal of integrative agriculture, 18(1), pp.54-61.