Predicting and forecasting of time series models using cluster analysis

I Setiawan¹*, I M Sumertajaya² and F M Afendi²

¹ Statistics Studies Program, Tadulako University, Palu, Indonesia
² Department of Statistics, IPB University, Bogor Indonesia

*Email: i.setiawan@untad.ac.id

Abstract. Time series model on multiple objects could be univariate and multivariate model. The more objects used in multivariate model would decrease precision of forecast for each object. One way to overcome this problem used univariate models for each object. However, univariate models for each object became inefficient in time. Therefore, clustering performed on objects so that model became efficient. The objective of this research is to study results of predicting and forecasting model with and without clustering. Model and forecasting used time series regression model on broad proportion of plant-disturbing organism attack and planting area of food crops in Indonesia. Clustering of time series data was same as clustering in general, but the distance and method should be able to accommodate time series data structure which was dynamic in time. Evaluation of prediction and forecast mean average percentage error (MAPE) show that forecast of model was performed with clustering as good as forecast of model for each object. However, prediction of model with clustering was not as good as the prediction of model for each object so that the prediction of broad proportion of plant-disturbing organism attack only served as an indicator of the arrival populations of plant-disturbing organism.

1. Introduction

Time series models could be approached using univariate or multivariate methods. Multivariate model is performed when there is a linear relationship between variables. However, the more variables used in multivariate model could decrease the precision of forecast for each variable. One way to overcome this problem is to perform univariate model for each variable, but this model could be inefficient in time. In order to obtain an efficient model, the variables are clustered.

Clustering of time series data is the same as clustering in general, however the selection of distance and the method of clustering used must be able to accommodate time series data structures that are dynamic with time. Liao classified the distances in time series clustering into three groups; raw data, feature and model [4]. Raw data is used directly with the raw data, feature is the distance representation of the data characteristics and model is the distance coefficient of the model. The example of time series using feature is Autocorrelation Function (ACF) by [1] and the example of time series using distance between model coefficients is Piccolo Distance (PIC) by [7]. The clustering methods are divided into hierarchical and non-hierarchical methods. Hierarchical method is a clustering with an unknown number of groups, meanwhile non-hierarchical method is a clustering with information on the number of groups known. Many researches related to time series clustering have been carried out. Jha et al. conducted a clustering to forecast sparse time series data [2].
The infested area of the main plant pest organism of food crop commodities is one of the time series data. Modelling and forecasting need to be carried out considering the impact given by the main plant pest infestation which is a decrease in the quality and quantity of food crop productions. We used the proportion of the main plant pest infestation on the planting area of food crops in order to obtain information on the actual attack rate of the main plant pests. Clustering was performed on the main plant pests of food crops that have the same pattern of the proportion of the infested area. However, there was no information on the number of clusters based on the proportion of available infested area. There are several studies related to the prediction of the main plant pests of food crops. [6] forecasted rice stem borer using climate factor information. The results obtained indicated that the forecast tended to be greater than the observation data.

In this study, we modelled and forecasted the proportion of the main plant pest infestation for food crops using the univariate modelling with and without clustering. We clustered the main plant pests of food crops using the agglomerative hierarchical clustering with Euclidean distance, PIC and ACF, then we modelled and forecasted each cluster. The purpose of this study is to evaluate the forecasting models with and without clustering on the proportion of the main plant pest infestation for food crops in Indonesia.

2. Materials and methods

2.1. Materials

The data used in this study obtained from Directorate General of Crops is the proportion of the main plant pest infestation on the planting area in commodity rice (Oryza sativa), corn (Zea mays), soybeans (Glycine max), peanuts (Arachis hypogaea), green beans (Vigna radiata), cassava (Manihot esculenta) and sweet potatoe (Ipomoea batatas). And the proportion of flood and drought areas. The explanatory variables used are physical environmental factors (climate factors) sourced from the Meteorological, Climatological and Geophysical Agency. The time period for data is the monthly period from January 2010 - December 2016. The data is divided into two parts, namely January 2010 - December 2015 as training data and January 2016 - December 2016 as testing data.

Table 1. Variables used in the study

| Variable | Information                              | Information                              |
|----------|------------------------------------------|------------------------------------------|
| Y₁       | Proportion of the main plant pest infestation | X₆                                       |
| X₁       | Rainfall (mm)                             | X₇                                       |
| X₂       | Humidity (%)                             | X₈                                       |
| X₃       | Average Temperature (°C)                 | X₉                                       |
| X₄       | Average Wind Speed (knot)                | X₁₀                                      |
| X₅       | Long Exposure (hour)                      |                                          |

Table 2. Plant pest organism used in the study

| Variable | Plant Pest Organism   | Variable | Plant Pest Organism   |
|----------|-----------------------|----------|-----------------------|
| Y₁       | Rice stem borer       | Y₁₄      | Soybean leaf rollers  |
| Y₂       | Planthopper brown rice| Y₁₅      | Soybean fly           |
| Y₃       | Rice rat              | Y₁₆      | Soybean pod borer     |
| Y₄       | Rice blast            | Y₁₇      | Soybean caterpillar   |
| Y₅       | Rice BLB              | Y₁₈      | Peanut brown leaf spots|
| Y₆       | Rice tungro           | Y₁₉      | Peanut mice           |
| Y₇       | Corn cobs borer       | Y₂₀      | Peanut leaf rust      |
| Y₈       | Corn stem borer       | Y₂₁      | Mung bean mice        |
| Y₉       | Corn armyworm         | Y₂₂      | Cassava wild boar     |
| Y₁₀      | Corn seed fly         | Y₂₃      | Cassava brown leaf spots|
| Y₁₁      | Cornstarch            | Y₂₄      | Cassava mice          |
| Y₁₂      | Corn rat              | Y₂₅      | Sweet potato wild boar |
| Y₁₃      | Soybean armyworm      |          |                       |
2.2. Methods

The analysis used in this study is divided into several stages as follows:

1. Dividing the data into training data for modelling and testing data for forecasting.

2. Modelling the training data using time series regression model with logit transformation on response variable \( Y \). The steps in modelling are as follows:
   a. Identifying the stationary assumption on the mean using ADF test. If the assumption is not met, differencing (d and D) is performed.
   b. Performing the regression model using response variable \( (Y) \) and explanatory variables \( (X) \) in Table 1 with the residuals following ARIMA \((2,d,0)\) on non-seasonal data and SARIMA \((2,d,0)\) \((1,D,0)\)m on seasonal data.
   c. Performing ARIMA or SARIMA models on the residuals of the regression model. The steps are as follows:
      i. Selecting the initial model according to the smallest AICc value from the model:
         - Non-seasonal: ARIMA \((2,d,2)\), ARIMA \((0,d,0)\), ARIMA \((1,d,0)\) and ARIMA \((0,d,1)\).
         - Seasonal: SARIMA \((2,d,2)\) \((1,D,1)\)m, SARIMA \((0,d,0)\) \((0,D,0)\)m, SARIMA \((1,d,0)\) \((1,D,0)\)m and SARIMA \((0,d,1)\) \((0,D,1)\)m.
      ii. Selecting the new model according to the smallest AICc value by simulating the order of the initial model:
         - Simulation 1: adding one of \( p, q, P \) and \( Q \) with \( \pm 1 \).
         - Simulation 2: adding \( p \) and \( q \) with \( \pm 1 \).
         - Simulation 3: adding \( P \) and \( Q \) with \( \pm 1 \).
         - Simulation 4: with or without constant.
      iii. Repeating point (ii) on the new model. This process stops if the new model obtained has an AICc value greater than the previous model.
   d. Performing the regression model with the residuals following the order of ARIMA or SARIMA models at point c.
   e. Performing diagnostic models on residuals and evaluating the prediction of the proportion of infested area in 2010-2015 using training data and
      \[
      \text{MAPE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i| 	imes 100}{\sum_{i=1}^{n} |y_i|} \times 100 \%
      \]  
   f. Repeating points a - e using a combination of explanatory variables \((X)\) in Table 1.
   g. Selecting the optimal combination of explanatory variables based on the smallest MAPE value.
   h. Modelling points a - e using the optimal combination of explanatory variables.
   i. Forecasting the testing data for the next 12 months and evaluating the results using MAPE.

3. Modelling and forecasting the training data using agglomerative hierarchical clustering with complete linkage.
   a. Calculating the distance matrix
      i. Euclidian distance is calculated using the proportion of the main plant pest infestation for food crops. [3]
      \[
      d_{eucl}(x, y) = ((x - y)'(x - y))^{1/2}
      \]
      ii. ACF distance is calculated using autocorrelation Euclidean distance \((\hat{p})\) at lag 1-71. [1]
      \[
      d_{ACF}(x, y) = ((\hat{p}_x - \hat{p}_y)'(\hat{p}_x - \hat{p}_y))^{1/2}
      \]
      iii. PIC distance is calculated using the autoregression coefficient of Euclidean distance \((\hat{\rho})\). [7]
      \[
      d_{PIC}(x, y) = ((\hat{\rho}_x - \hat{\rho}_y)'(\hat{\rho}_x - \hat{\rho}_y))^{1/2}
      \]
b. Choosing the optimal number of clusters using the ratio of the total variance between and within groups with the value of the F distribution and choosing the optimal distance using the total variance within small groups.

c. Performing clustering using the optimal number of clusters and distances.

d. Choosing cluster prototypes. Given cluster \( \mathcal{C} = \{x_1, x_2, ..., x_c\} \) sized \( N \) then the median of the \( i \)-th cluster is \( \text{prototype}_i = \text{median} \{x_{1,i}, x_{2,i}, ..., x_{c,i}\} \) \( \forall c \in \mathcal{C} \) [8]

e. Modelling and forecasting the prototypes as in point 2.

f. Evaluating the prediction and the forecast of prototype on the proportion of the main plant pest infestation for food crops in the corresponding cluster using MAPE.

4. Comparing the model and the forecast with and without clustering using MAPE and the average of MAPE.

3. Results and discussion

3.1. Modelling and forecasting training data without clustering

We used time series regression model using physical environmental factors (\( X \)) and response variable (\( Y \)) which had been log transformed as shown in Table 1. Given an illustration, a model of rice blast where the residuals of regression model follows SARIMA (0,1,2) (0,1,1)\( \times 12 \).

\[
\ln \left( \frac{Y_t}{1 - Y_t} \right) = -0.032X_2 + 0.605X_3 - 0.711X_4 - 0.251X_5 - 0.507X_6 - 0.176X_7 + 3.502X_9 + 1.146X_{10} + \varepsilon_{t-1} + \varepsilon_{t-12} - w_t + 0.420w_{t-1} + 0.492w_{t-2} + 0.634w_{t-12} + 0.266w_{t-13} + 0.311w_{t-14}
\]

The model coefficient value shows the strength of the linear relationship between physical environmental factors and the ratio of the proportions of the infested and non-infested areas. For example, the interpretation of the effect of average wind speed (\( X_4 \)) with other physical environmental factors which are assumed to be constant, is if the average wind speed which increases by one knot, then the ratio of the proportion of rice blast infested and non-infested areas increases by \( e^{-0.711} = 0.491 \) times.

3.2. Clustering

We performed agglomerative hierarchical clustering with complete linkage to obtain the optimal number of clusters and the optimal distance.

![Figure 1](image-url)

Figure 1. Evaluations of (a) the ratio of total variance between and within groups and (b) total variance for the optimal number of clusters.

Figure 1 (a) shows that the optimal number of clusters using Euclidian distance is four with a variance ratio of 3.23, the optimal number of clusters using piccolo distance is seven with a variance ratio of 2.67 and the optimal number of clusters using ACF distance is four with a variance ratio of 2.87. Figure 1 (b) shows that the clustering using Euclidian distance at the optimal cluster obtains the similarity pattern of
the infestation proportion in one group which is better than using other distances. Hence, Table 3 shows the results of clustering with Euclidian distance performed in the next analysis.

Table 3. Results of clustering with Euclidean distance

| Cluster | Plant Pest Organism |
|---------|---------------------|
| 1       | Rice rat, rice stem borer, rice BLB |
| 2       | Rice blast, rice planthopper, rice stem borer, corn stem borer, corn armyworms, corn seed fly, cornstarch, corn rat, soybean fly, peanut mice, mung bean mice, cassava wild boar, cassava brown leaf spots, cassava mice, sweet potatoes wild boar, soybean armyworms, peanut leaf rust, soybean leaf rollers, soybean caterpillar, soybean pod borer, |
| 3       | Peanut brown leaf spots |
| 4       | Peanut brown leaf spots |

Table 3 shows the same pattern of infestation proportions for each cluster. The proportion of infested area in Cluster 1 follows a seasonal pattern, in Cluster 2, the proportion follows a seasonal pattern and contains outliers. Cluster 3 is a group with an infestation data pattern in 2010-2011 which is smaller compared the other years, with high local variations following the time. Cluster 4 is relatively the same as Cluster 3 and contains outliers.

3.3. Modelling with clustering

First, we chose the prototype of each group clustered by complete linkage with Euclidian distance. Figure 2 shows the illustration of Cluster 1 prototype.

Figure 2 shows that the cluster prototype sufficiently represents the pattern of the proportion of the main plant pest infestation in Cluster 1. The Cluster 1 prototype has a pattern similar to the pattern of the proportion of rice stem borer infestation. We modelled and forecasted Cluster 1 prototype using time series regression model with physical environmental factors (X) and the response variable (Y) which had been log-transformed.

Table 4 shows that the structure of the prototype modelling in Cluster 1 with the residuals following SARIMA (2,1,1) (0,1,1)_{12} is similar to the structure of the rice stem borer model with the residuals following SARIMA (0,0,2) (0,1,1)_{12}.
The 2-nd International Seminar on Science and Technology 2020 (ISST-2) 2020
Journal of Physics: Conference Series 1763 (2021) 012035 doi:10.1088/1742-6596/1763/1/012035

Table 4. The structure of the prototype modelling in Cluster 1

| Plant Pest Organism | Without Clustering | With Clustering |
|---------------------|-------------------|----------------|
|                     | $\ln \left( \frac{Y_t}{1 - y_t} \right)$ | $\ln \left( \frac{Y_t}{1 - y_t} \right)$ |
| Rice stem borer     | $-0.009 - 0.006X_1 - 0.044X_2 +$ | $-0.036X_1 - 0.018X_2 -$ |
|                     | $0.067X_3 - 0.255X_4 - 0.130X_5 -$ | $0.139X_3 - 0.107X_4 -$ |
|                     | $0.027X_6 + 1.635X_8 + 0.434X_{10} + \varepsilon_{t-12} -$ | $0.140X_5 - 0.206X_9 +$ |
|                     | $w_t + 0.038w_{t-1} + 0.513w_{t-2} +$ | $0.379X_8 + 0.978\varepsilon_{t-1} -$ |
|                     | $0.630w_{t-12} + 0.024w_{t-13} + 0.323w_{t-14}$ | $0.622\varepsilon_{t-2} + 0.622\varepsilon_{t-14} +$ |
| Rice rat            | $-0.028X_1 + 0.028X_2 + 0.112X_8 +$ | $0.363\varepsilon_{t-3} + 0.643\varepsilon_{t-15} + w_t +$ |
|                     | $1.998X_9 + 0.916\varepsilon_{t-1} - 0.792\varepsilon_{t-2} +$ | $0.485w_{t-1} + 0.726w_{t-12} +$ |
|                     | $0.629\varepsilon_{t-12} - 0.576\varepsilon_{t-13} + 0.498\varepsilon_{t-14} +$ | $0.352w_{t-13}$ |
|                     | $0.371\varepsilon_{t-24} - 0.340\varepsilon_{t-25} + 0.294\varepsilon_{t-26} +$ | |
|                     | $0.876\varepsilon_{t-3} - 0.550\varepsilon_{t-15} - 0.325\varepsilon_{t-27} +$ | |
|                     | $w_t + 0.823w_{t-12}$ | |
| Rice BLB            | $-0.065X_2 - 0.579X_4 - 0.219X_5 -$ | |
|                     | $0.423X_6 + 0.305X_7 + 0.154X_9 -$ | |
|                     | $1.769X_8 + 1.344\varepsilon_{t-1} - 0.948\varepsilon_{t-2} +$ | |
|                     | $\varepsilon_{t-12} - 1.344\varepsilon_{t-13} + 0.948\varepsilon_{t-14} +$ | |
|                     | $0.605\varepsilon_{t-3} - 0.605\varepsilon_{t-15} + w_t +$ | |
|                     | $0.646w_{t-12}$ | |

3.4. The comparison of modelling with and without clustering

We compared the prediction of the models using the average of MAPE as shown in Table 5.

Table 5. The average of MAPE for prediction and forecast

| Evaluation | Without clustering (%) | With Clustering (%) | t-value | p-value |
|------------|------------------------|---------------------|---------|---------|
| Prediction | 49.961                 | 152.628             | -2.702  | *0.010  |
| Forecast   | 83.108                 | 103.52              | -0.945  | 0.349   |

*Significant at 5 % level of significance

Table 5 shows that the difference in the predictions of modelling with and without clustering is quite huge. On the other hand, the difference in the forecasting models is quite small. The t test shows that the predictions for both models are significantly different, meanwhile the forecasts for both models are relatively equal. Figure 3 shows the comparison of modelling with and without clustering.

Figure 3 (a) shows that modelling with clustering the data first obtains a MAPE value greater than the modelling without clustering. This is due to the fact that the prediction does not conform to a fairly high pattern of the proportion of the infested area. However, Figure 3 (b) shows that the modelling with clustering obtains a MAPE value relatively equal with the modelling without clustering. The evaluation shows that the average forecast for both models are relatively equal, hence the efficiency of the clustering process.

The evaluations which are not good enough indicate that the prediction and forecast obtained only serve as indicators of the population of the main plant pests of food crops in the planting area causing the infestation, therefore these cannot be able to explain the actual proportion of the infested area. On the other hand, the good evaluations indicate that the prediction and forecast obtained do not only serve as indicators of the said population but also are able to explain the actual proportion of the infested area.
4. Conclusion

Based on the discussion, we conclude that the model with clustering provides efficiency in the process because the forecast obtained is relatively equal with the model without clustering. However, the prediction of the infested area is not as good, therefore, it only serves as an indicator of the population of the main plant pests of food crops in the planting area.

Acknowledgement

Thank you to the Head of the Faculty of Mathematics and Natural Sciences, University of Tadulako for funding assistance so that this research can be carried out.

References

[1] D'Urso P and Maharaj E A 2009 Fuzzy Sets and Systems 160 3565-3589
[2] Jha A Ray S Seaman B and Dhillon I S 2015 2015 IEEE 31st International Conf. on Data Engineering Seoul 388-1399
[3] Johnson R A and Wichern D W 2007 Applied Multivariate Statistical Analysis 6th Edition (New Jersey (US): Prentice Hall Inc)
[4] Liao T W 2005 Pattern Recognition 38 1857-1874
[5] Montgomery D C Jennings L C and Kulahci M 2008 Introduction to Time series Analysis and Forecasting (New Jersey (US): John Wiley & Sons Inc)
[6] Nurhayati E 2017 Pengembangan sistem informasi iklim berbasis pemodelan dinamika populasi untuk rencana pengendalian penggerek batang padi kuning Bogor (ID): Institut Pertanian Bogor
[7] Piccolo D 1990 Journal of Time Series Analysis 11(2) 153-164
[8] Sarda-Espinosa A 2019 The R Journal 11(1) 22-43