A Technique of Fuzzy C-Mean in Multiple Linear Regression Model toward Paddy Yield

Nur Syazwan Wahab, Mohd Saifullah Rusiman, Mahathir Mohamad, Nur Amira Azmi, Norziha Che Him, M.Ghazali Kamardan and Maselan Ali

Fakulti Sains, Teknologi dan Pembangunan Insan, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia

E-mail: nursyazwanwahab93@gmail.com, saifulah@uthm.edu.my, mahathir@uthm.edu.my

Abstract. In this paper, we propose a hybrid model which is a combination of multiple linear regression model and fuzzy c-means method. This research involved a relationship between 20 variates of the top soil that are analyzed prior to planting of paddy yields at standard fertilizer rates. Data used were from the multi-location trials for rice carried out by MARDI at major paddy granary in Peninsular Malaysia during the period from 2009 to 2012. Missing observations were estimated using mean estimation techniques. The data were analyzed using multiple linear regression model and a combination of multiple linear regression model and fuzzy c-means method. Analysis of normality and multicollinearity indicate that the data is normally scattered without multicollinearity among independent variables. Analysis of fuzzy c-means cluster the yield of paddy into two clusters before the multiple linear regression model can be used. The comparison between two method indicate that the hybrid of multiple linear regression model and fuzzy c-means method outperform the multiple linear regression model with lower value of mean square error.

1. Introduction

Paddy is the third most widely planted crop in Malaysia after oil palm and rubber. Based on the Paddy Statistics of Malaysia, in 2000, 698,710 hectares were planted with paddy. In 2010, 687,940 hectares were planted with paddy, whereas in 2014, 679,239 hectares were planted with paddy including those that are planted twice a year. Paddy rice is important in the paddy industry for planning and development of the industry. Rice is the staple food (main) populace in Malaysia. Therefore, the positive efforts do to fulfil the domestic needs [1, 2]. The Ninth Malaysia Plan (9MP) which was launched by the government in 2006, the Malaysian government began to move and focus the progress of agricultural sector towards the advanced and sustainable of agricultural systems. Ongoing transformation requires acculturation in terms of the use of technology among farmers in Malaysia [3].

Paddy is an important crop in Malaysia and it is vital for the nation's food security. Apart from this, the statistic also has proven that paddy industry in Malaysia has generated stable income for the country. Such income generation has reflected the success of this industry. Nonetheless, is the success of this industry has any impact on the paddy farmers particularly on their well-being. This query has become the main objective of this paper which is to discover the impingement factors of paddy farmers’ well-being. This is qualitative study where data were gained from documents and literature analyses. Based on the analyses performed, it can be seen that factors such as financial, social and human should be considered to further enhance the farmers’ well-being [4].

Alam et al. [5] studied the changing nature of climatic factors has different impact on agriculture based on areas, time periods, and crops. Farmers are the most vulnerable group who are affected both
directly and indirectly by climatic changes. This study focuses on the impact of climatic changes in Malaysia, specifically in the Integrated Agricultural Development Area (IADA), in Northwest Selangor. Due to changes in the climate, the productivity and profitability of paddy cultivation have declined in this area. Carmen [6] concluded that the paddy response to applied fertilizers varies under different soil, climatic and management conditions. Rates of fertilizers which are adequate for a certain yield level in one location could not be the same in another location with different environmental conditions.

Regression analysis was first developed by Sir Francis Galton in the latter part of the 19th century. Regression analysis has become one of the standard tools in data analysis. Its popularity comes from different disciplines. The mathematical equation from its analysis could explain the relationship between the dependent and independent variables. It has been widely used in applied sciences, economic, engineering, computer, social sciences and other fields. Multiple regression is a flexible method of data analysis that may be appropriate whenever a quantitative variable is to be examined in relationship to any other factors [7].

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is developed by Dunn [8] and improved by Bezdek [9]. It is frequently used in pattern recognition. Especially, fuzzy clustering has been widely studied and applied in a variety of key areas and fuzzy cluster validation plays a very important role in fuzzy clustering. There are other quite considerable studies were carried out to merge MLR model with other method nowadays such as fuzzy theory [10, 11, 12, 13]. Whereas [14, 15, 16] studied the statistical application used in Malaysia.

This study is about the multiple linear regression model and fuzzy c-means in multiple linear regression technique toward paddy yield. The purpose of this study is to investigate the factors that affect the paddy yield. Through this study the investigation of how variables involved and then giving impact toward paddy yield can be found. Lastly, the model from both multiple linear regression analysis and fuzzy c-means technique analysis will be compared by using the smaller value of mean square error in order to find the better model.

2. Methodology

There are 20 independent variables that are analysed using least square method toward paddy yield before adding fertilizer. The independent variables sample (x1), sand (x2), silt (x3), clay (x4), PH of water (x5), PH of potassium chloride (x6), electric conductors (x7), solution phosphorus (x8), organic carbon (x9), nitrogen (x10), change of positive charge (x11), changes in potassium (x12), changes in sodium (x13), changes in calcium (x14), changes in magnesium (x15), changes in aluminium (x16), manganese (x17), separation of potassium (x18), separation of sodium (x19) and absorption of alkali (x20). While for dependent variable is y (paddy yield in kg/hectare). The model of this study should be able to provide an ideal relationship between all independent variables with predictor variables. From this study also, it should be able to identify the factors that give more impact in data of paddy yield.

In this study, there are two analysis used which are multiple linear regression analysis and multiple linear regression with fuzzy c-means technique. By using SPSS software and R packages, test of normality and multicollinearity produced. Firstly, normality test is done by using Shapiro-Wilk and Kolmogorov Smirnov test. Next, the multicollinearity test is done using VIFF test. If VIFF value is less than 10 for each X variables, it indicates that multicollinearity between X variables do not exist [17].

Multiple regression models is a linear regression model with two or more predictors and one response. There are also regression models with two or more response variables. These models are usually called multivariate regression models. Here, this will introduce a new (linear algebra based) method for computing the parameter estimates of multiple regression models. This more compact method is convenient for models for which the number of unknown parameters is large. For example, a multiple linear regression model with k predictor variables X_1, X_2,...,X_k and a response Y can be written as in (1).

\[ Y = \beta + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon. \]  

(1)

As before, the \( \epsilon \) are the residual terms of the model and the distribution assumption we placed on the residuals will allow us later to do inference on the remaining model parameters. Interpret the meaning of the regression coefficients \( \beta_0, \beta_1, \beta_2 \ldots \beta_k \) in this model [17].
Models with two predictor variables (say $x_1$ and $x_2$) and a response variable $y$ can be understood as a two-dimensional surface in space. The shape of this surface depends on the structure of the model. The observations are points in space and the surface is “fitted” to best approximate the observations. For example, the simplest multiple regression model for two predictor variables as in (2).

$$y = \beta + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad (2)$$

Then, Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is developed by Dunn [8] and improved by Bezdek [9]. It is frequently used in pattern recognition. A large family of fuzzy clustering algorithms is based on minimization of the fuzzy c-means functional formulated as in (3).

$$J(Z; U, V) = \sum_{c}^{c} \sum_{j=1}^{N} u_{ij}^w d_{ij}^2 \quad (3)$$

where,

- $N$: The number of objects or observations
- $c$: The number of clusters.

Before the formation of MLR model of fuzzy c-means, the data is clustered into $c$ group of clustering by using Matlab software. Next, the model of MLR is produced from $c$ group of clustering by using SPSS. The MSE for both model with applied fuzzy c-means technique being obtained. Finally, all the MSE value of the model will be compare in order to find the best model and factors for the paddy yield.

### 3. Result and discussions

There are missing observations for some variables. Missing observations is overcome by using the mean of variable. In this study, the normality test was performed to see if the residuals are normally distributed by using Shapiro-Wilk normality test and Kolmogorov Smirnov test as in Table 1. Both of the tests show that the residual are normally distributed since the $p$-value is bigger than 0.05. It can be concluded that the data are normally distributed.

**Table 1. P-value of normality test**

| Type of test                              | P-value |
|------------------------------------------|---------|
| Kolmogorov-Smirnov Normality Test        | 0.9896  |
| Shapiro-Wilk Normality Test              | 0.5514  |

Besides that, multicollinearity is tested by correlating the explanatory variables against each other. Through an eigenvalues of paddy yield model, it shows no eigenvalues is smallest compared with the others eigenvalues. Next, the VIFF test is less than 10 for each $X$ variables which indicate that multicollinearity between $X$ variables do not exist. So, the model of paddy yield shows that there is no multicollinearity.

**Table 2. ANOVA table for multiple linear regression model**

|                | Sum of Squares | df  | Mean Square   | F          | Sig.   |
|----------------|----------------|-----|---------------|------------|--------|
| Regression     | 178573221.146  | 20  | 8928661.057   | 48.075     | .000   |
| Residual       | 28044298.854   | 151 | 185723.833    |            |        |
| Total          | 206617520.000  | 171 |               |            |        |
Refer to the ANOVA table (analysis of variance) in Table 2 above, it was found that the value of MSE is 185,723.833. A model of paddy yield before adding fertilizer is significant at the 5% confidence level and the null hypothesis is rejected. In other word, the model of MLR is significant as in (5). The hypothesis for this study is in (4).

\[ H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = \beta_{17} = \beta_{18} = \beta_{19} = \beta_{20} = 0 \]

\[ H_1 : \text{At least one } \beta_i \neq 0 \]  

(4)

\[ y = 18363.008 - 159.52x_1 - 123546x_2 - 131.144x_3 - 249.741x_4 + 293.45x_5 - 2062.7x_6 + 0.908x_7 - 47.914x_8 + 3418.83x_9 - 188.216x_{10} - 854.232x_{11} + 734.859x_{12} - 405.43x_{13} + 47.789x_{14} + 40.49x_{15} - 23.755x_{16} - 292.055x_{17} + 419.01x_{18} + 158.145x_{19} - 65.812x_{20} \]  

(5)

Next, starting with the fuzzy c-means (FCM) method, a set of 172 data patterns was considered. The objective is to cluster data into ‘c’ clusters by minimizing the objective function. Using fuzzy c-means method, it is found that the suitable cluster is 2 using Matlab software. Next, the model of MLR is produced from 2 group of clustering by using SPSS software as shown in Figure 1 and Figure 2. Finally, all the MSE value of the model will be compare in order to find the best model and factors for the paddy yield.

For Cluster 1:  
\[ \frac{2813 - 2612}{a - 2612} = \frac{0.5673 - 0.3856}{0.4765 - 0.3856} \]
\[ a = 2712.555 \]

For Cluster 2:  
\[ \frac{2813 - 2612}{a - 2612} = \frac{0.4327 - 0.6144}{0.5236 - 0.6144} \]
\[ a = 2712.445 \]

Figure 1 shows that the paddy yield is ranging from 1212 to 2712 kg/hectare for Cluster 1, whereas the paddy yield is ranging from 2713 to 5163 kg/hectare for Cluster 2. Besides, Figure 2 shows the membership value for Cluster 1 and Cluster 2.
Figure 2. The fuzzy c-mean (FCM) graph of membership function with two clusters

For FCM technique in MLR, the Stepwise method is used to analyse the MLR model. From each Cluster 1 and Cluster 2 of fuzzy c-means, the model of $y$ for each clusters can be formed. The MLR model of $y$ for each cluster is shown in Table 3 below. For Cluster 1, there are 12 independent significant variables, whereas for Cluster 2, there are only 7 independent significant variables.

Table 3. Summary model of FCM in MLR model

|                      | Cluster 1               | Cluster 2          |
|----------------------|-------------------------|--------------------|
|                      | (From data 1 to 52)     | (From data 53 to 172) |
| MLR model            | $y(hat) = 5263.507 - 51.847X_2$ | $y(hat) = 3891.562 - 23.215X_1$ |
|                      | $-36.428X_3 + 391.817X_4 - 384.284X_5$ | $+23.976X_2 + 205.092X_5$ |
|                      | $+211.783X_6 + 38.129X_{10}$ | $+726.058X_6 + 100.049X_8$ |
|                      | $+1537.486X_{11} - 959.066X_{13}$ | $-75.358X_{10} - 161.160X_{13}$ |
|                      | $-177.513X_{15} + 317.433X_{17}$ |                      |
|                      | $+365.333X_{18} + 9.391X_{20}$ |                      |
| MSE                  | 0.000                   | 92074.877         |

Then, the value of mean square error (MSE) also can be found after build the model of MLR for both cluster of model. This study concluded that the multiple linear regression technique using fuzzy c-means (FCM) is the better model if compared to the multiple linear regression without using fuzzy c-mean (FCM) by referring to the smaller value of MSE. In fact, the value of MSE for multiple linear regression without using FCM technique is 185,723.833, whereas the value of MSE for multiple linear regression with using FCM technique is 0 (Cluster 1) and 92074.877 (Cluster 2).

4. Conclusion

In conclusion, according to the multiple linear regression and fuzzy c-means, there are missing observations for some variables. Missing observations is overcome by using the mean of variable. There are many independent variables involved in this study and the insignificant variables are discarded. Analysis showed that in this study there was no problem in the normality and multicollinearity. The
analysis also showed that the paddy yield is direct proportional to sand \((x_2)\), PH of water \((x_5)\), PH of potassium chloride \((x_6)\) and solution phosphorus \((x_8)\), whereas the paddy yield is inversely proportional to nitrogen \((x_{10})\) and changes in sodium \((x_{13})\). Finally, the paddy yield model obtained in this study is very useful in estimating or predicting the mean value for yield potential of rice or rice production in a particular place before adding fertilizer. However, differences in other factors such as climate, crop management and others are assumed a little bit different. If the paddy yield in a particular place is less, positive steps should be taken to increase paddy yield.

Acknowledgement

This research work is supported by FRGS (Fundamental Research Grant Scheme) grant (Vot 1498), Ministry of Higher Education, Malaysia.

Reference

[1] Jabatan Pertanian Semenanjung Malaysia 2012 Perangkaan Padi Malaysia (Paddy Statistics Of Malaysia) 2011 (Jabatan Pertanian Semenanjung Malaysia)

[2] Jabatan Pertanian Semenanjung Malaysia 2015 Perangkaan Padi Malaysia (Paddy Statistics Of Malaysia) 2014 (Jabatan Pertanian Semenanjung Malaysia)

[3] Oxford Business Group 2007 The Report Malaysia 2007 (Oxford Business Group)

[4] Fahmi Z, Samah B A and Abdullah H 2013 Paddy Industry and Paddy Farmers Well-being: A Success Recipe for Agriculture Industry in Malaysia Asian Social Science 9 (3) 177-181

[5] Alam M M, Talib B, Siwar C, Toriman M E 2010 The Impacts of Climate Change on Paddy Production in Malaysia: Case of Paddy Farming in North-West Selangor. Proceedings of the international conference of the 4th International Malaysia - Thailand Conference on South Asian Studies. National University of Malaysia, Malaysia 25-26

[6] Carmen M L 1968 Yield of Rice as Affected by Fertilizer Rates, Soil and Meteorological Factors (Iowa State University)

[7] Agresti A 1996 An Introduction to Categorical Data Analysis (New York: John Wiley and Sons)

[8] Dunn J C 1973 A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters Journal of Cybernetics 3 32-57

[9] Bezdek J C 1981 Pattern Recognition with Fuzzy Objective Function Algorithms (USA: Kluwer Academic Publishers)

[10] Bin Shafi M A, Bin Rusiman M S and Che Yusof N S H 2014 Determinants Status of Patient After Receiving Treatment at Intensive Care Unit: A Case Study in Johor Bahru. I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings 30 September 2014, 6914150 80 – 82

[11] Rusiman M S, Nasibov E and Adnan R 2011 The Optimal Fuzzy C-regression Models (OFCRM) in Miles per Gallon of Cars Prediction, Proceedings – 2011 IEEE Student Conference on Research and Development, SCOReD 2011, 6148760 333-338

[12] Shafi M A and Rusiman M S 2015 The Use of Fuzzy Linear Regression Models for Tumor Size in Colorectal Cancer in Hospital of Malaysia Applied Mathematical Sciences 9 (56) 2749-2759

[13] Kavikumar J, Khamis A and Rusiman M S 2013 N-Structures Applied to Finite State Machines, IAENG International Journal of Applied Mathematics 43 (4) 233 – 237

[14] Sufahani S, Che-Him N, Khamis, A, Rusiman M S, Arbin N, Yee C K, Ramli I N, Suhaimi N A, Jing S S and Azmi Z A 2017 Descriptive Statistics with Box-Jenkins and Marketing Research for Jewellery Company in Malaysia Far East Journal of Mathematical Sciences 101(10) 2151-2161

[15] Nor M E, Khamis A, Saharan S, Abdullah M A A, Salleh R M, Asrah N M, Khalid K, Aman F, Rusiman M S, Halim H, Lee M H and Nor E 2016 Malaysia Tourism Demand Forecasting by Using Time Series Approaches Social Science (Pakistan) 11(12) 2938-2945
[16] Nor M E, Rusiman M S, Mohamad N A I and Lee M H 2017 Directional Change Error Evaluation in Time Series Forecasting AIP Conference Proceedings 1830 (1) 080013
[17] Montgomery D C and Peck E A 1992 Introduction to Linear Regression Analysis (2nd Edition) (New York: John Wiley & Sons)