International Journal of Engineering & Technology, 7 (4.30) (2018) 473-478

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A new hybrid of Fuzzy C-Means Method and Fuzzy Linear Regression Model in Predicting Manufacturing Income

Nurfarawahida Ramly[1], Mohd Saifullah Rusiman*[1], Norziha Che Him*[1], Maria Elena Nor*[1], Suparman[2], NurAin Zafirah Ahmad Basri[1], Nazeera Mohamad[1]

[1]Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Pagoh Educational Hub, 84600 Pagoh, Muar, Johor, Malaysia
[2]Department of Mathematics Education, University of Ahmad Dahlan, Indonesia
**Corresponding author E-mail: saifulah@uthm.edu.my*

Abstract

Analysis by human perception could not be solved using traditional method since uncertainty within the data have to be dealt with first. Thus, fuzzy structure system is considered. The objectives of this study are to determine suitable cluster by using fuzzy c-means (FCM) method, to apply existing methods such as multiple linear regression (MLR) and fuzzy linear regression (FLR) as proposed by Tanaka and Ni and to improve the FCM method and FLR model proposed by Zolfaghari to predict manufacturing income. This study focused on FLR which is suitable for ambiguous data in modelling. Clustering is used to cluster or group the data according to its similarity where FCM is the best method. The performance of models will measure by using the mean square error (MSE), the mean absolute error (MAE) and the mean absolute percentage error (MAPE). Results shows that the improvisation of FCM method and FLR model obtained the lowest value of error measurement with MSE=1.825×10^{11} , MAE=115932.702 and MAPE=95.0366. Therefore, as the conclusion, a new hybrid of FCM method and FLR model are the best model for predicting manufacturing income compared to the other models.

Keywords: *Fuzzy linear regression (FLR), fuzzy c-means (FCM), mean square error (MSE)*

1. Introduction

Cluster analysis is the art of finding groups in datasets. The classification of similar objects into groups is an important human activity. In everyday life, this classification part is always used as learning process [1]. Besides, classification also appears in many disciplines such biology, medicine, psychology, marketing and image processing [2]. The choice of a clustering algorithm depends both on the type of data available and specific purpose. In clustering the data, two most widely studied clustering algorithms are partitional and hierarchical clustering [3].

These algorithms have been heavily used in wide range of applications primarily to their simplicity and to ease of implementation relative to other clustering algorithms. Partitional clustering algorithms aim to discover the groupings present in the data by optimising a specific objective function and iteratively improving the quality of the partitions. Partitional methods generally required a user predefined parameter to obtain clustering solution. Meanwhile, hierarchical clustering algorithms used the problem of clustering by developing a binary tree-based data structure. It was developed to build more deterministic and flexible mechanism for clustering the data objects [3]. Performing hard assignments of points to clusters is not feasible in complex datasets where there are overlapping clusters. To extract such overlapping structure, a FCM clustering algorithm could be used.

Fuzzy approach is successfully applied in various experiments which involved fuzzy data. Fuzzy regression is an important method for analysing vague association between response and explanatory variables. Fuzzy logic able to control complex systems more effectively compared to traditional approaches [4]. Here, Tanaka et al. [5] is the first proposing a fuzzy linear regression model which is useful for certain systems and significant to fuzzy structure and human estimation. Fuzzy linear regression could be categorised into two types of situations based on the functional relationship; dependent (response) and independent (explanatory) variables. Two types of categories are parametric fuzzy regression model where the functional relationship is known and nonparametric regression model if otherwise. Numerical method is used to identify the fuzzy regression model by minimising the sum of spreads of the estimated dependent variable. There are other quite considerable studies were carried out to use fuzzy and statistics techniques in Malaysia and other countries [17, 18, 19, 20, 21].

The fuzzy linear regression was focused on the FLR model with the assumption of triangular fuzzy numbers (TFNs) being either symmetrical or asymmetrical, where they both represents by its own membership function. The parameter in the FLR model could be estimated by certain methods. Finally, Zolfaghari et al. [6] considered two factors parameter estimation of fuzzy linear regression model, known as the degree of fitting and the vagueness of the model, which can transfer into two approaches.

2. Material and Method

The statistical software used in this study are MATLAB 2014, Microsoft Excel 2010 and SPSS 20. An exploratory data analysis was executed to explore the relationship between dependent variable and independent variables.

2.1. Multiple linear regressions

Multiple linear regression (MLR) analysis is an extension of simple linear regression (SLR) analysis [7]. MLR is among the commonly used statistical methods where it is highly useful in experimental situations of which the experimenter can control all predictor variables. MLR also attempts to develop any potential model on the relationship between two or more independent variables and dependent variable by fitting a linear equation to the observed data. The key assumptions of MLR model are linear association and no multicollinearity exist. The presence of linear association can be tested by using the Q-Q plot of the standardized predicted versus *Y* values. To avoid dependency amongst *X* variables, multicollinearity should be tested by using variance inflation factor (VIF) which means a measure of how much the variance of the estimated regression coefficient is inflated by the existence of correlation among the independent variables in the model. A VIF value of less than 10 means that there is less correlation or no correlation among the independent variable, meanwhile if the VIF value exceeds 10, it is the sign of serious multicollinearity and required some correction to be made [8].
This study used MLR to analyse nine independent variables that proof related to manufacturing income. This analysis was conducted to find the significant variable among the independent variables. In conducting the analyse need to consider the *p*-value with less than 0.05. Besides, the correlation coefficient value (r) and coefficient of determination value (r^2) also need to consider in constructing the MLR model. If x and y have a strong positive linear correlation, r is close to +1. Positive values indicated a relationship between x and y variables such that as values for x increases, values for y also increase. Meanwhile, if x and y shows a strong negative linear correlation, r is close to -1. Negative values indicated a relationship between x and y such that as values for x increase, values for y decrease. The MLR model such as in (1).

$$\hat{Y}_q = \beta_0 + \beta_1 X_{q1} + \beta_2 X_{q2} + ... + \beta_p X_{qp} + \varepsilon_q(\beta) \quad (1)$$

where $q = 1, 2, \ldots, N$, Y is the dependent variable with $X_1, ..., X_p$ are the independent variables and $\beta_1, ..., \beta_p$ is the regression coefficient. The function of least square method as in (2)

$$S(\beta_0, \beta_1, \beta_2, ..., \beta_p) = S(\beta) = \sum_{j=1}^{d} \varepsilon_j^2 \quad (2)$$

From (1), $\varepsilon(\beta) = Y - X\beta$, Thus

$$\begin{aligned} S(\beta) &= (Y - X\beta)^T (Y - X\beta) \\ &= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta \end{aligned} \quad (3)$$

In the least square model, the best fitting line for the observed data is calculated by minimising $S(\beta)$. Then, $S(\beta)$ will differentiate with respect to β where $\left.\frac{\delta S}{\delta \beta}\right|_{\beta}$ is equal to zero as in (4).

$$\left.\frac{\delta S}{\delta \beta}\right|_{\hat{\beta}} = -2X^T Y + 2X^T X \beta = 0 \quad (4)$$

Hence, the least square estimator is as shown in (5).

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (5)$$

The equation $\beta_0 + \beta_1 X_{q1} + \beta_2 X_{q2} + ... + \beta_p X_{qp} + \varepsilon_q$ is denoted as $\hat{Y}$ and the residual ε_q is equal to $Y_q - \hat{Y}_q$, which shows the difference between the observed and fitted values.

2.2. Fuzzy c-means

Fuzzy c-means (FCM) is a clustering method which allows one set of data belong to more than one cluster. [9] developed this method and later improved by [10]. Clustering is usually used as an alternative for segmentation techniques. This research applies clustering method to denote the techniques that been used in exploratory data analysis. Clustering methods attempt to group together patterns either that are similar or not similar in some sense. FCM-method has frequently used the pattern recognition where the algorithm is based on minimisation of FCM towards the following objective function or criterion in (6)

$$J = \sum_{q=1}^{N} \sum_{r=1}^{C} u_{qr}^z d_{qr}^2 \quad (6)$$

where *z* is any real number greater than 1, μ_{qr} is the membership values, d_{qr} represents the distance according to Euclidean, *N* is the number of objects and *C* is the number of clusters. The index *q* (*q* = 1,…,*N*) corresponds to object number *q* and the index *r* (*r* = 1,…, *C*) to cluster number *r*. In case of Euclidean distance, the algorithm for minimising *J* can be summarised as in the following steps:

1. Randomly select cluster centre, *c*. Choose the termination tolerance, δ between 0 and 1, and fuzziness exponent, $z > 1$ (usually $z = 2$). Fuzzy membership matrix, μ_{qr} is then initialised.

2. Update the distance d_{qr} for given μ_{qr} , independent variable, x_q and cluster center, v_r , by compute the weighted average for each group and the Euclidean distance as in (7).

$$d_{qr}^2 = \left\| x_q - v_r \right\|^2, \; v_r = \frac{\sum_{q=1}^{N} \mu_{qr}^z x_q}{\sum_{q=1}^{N} \mu_{qr}^z} \quad (7)$$

3. Update the membership values as in (8),

$$u_{qr} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{qr}}{d_{qk}} \right)^{\frac{2}{z-1}}}, \text{ for } z > 1 \quad (8)$$

$$u_{qr} = \begin{cases} 1 & \text{if } d_{qr} = \min(d_{qk}) \\ 0 & \text{otherwise} \end{cases}, \text{ for } z = 1 \quad (9)$$

4. Calculate the objective function or criterion, *J* in (6) and iterate to minimise the objective function. The iteration is repeated for $k = 1, 2, ..., \infty$. If $\left| J_{k+1} - J_k \right| < \delta$, stop the iteration, else repeat step 2 (δ is a termination error where usually 0.005 is chosen).

2.3. Cluster validity

Clustering results could be evaluated by clustering validation [11] which recognised as one of the vital issues essential to the success of clustering validation. There are two main categories of clustering validation which are the external and internal validation measure. This study focused on the internal validation measure which means it only rely on the information in the data and it evaluates the clustering structure without any external information [12]. Under internal validation measure of fuzzy c-means clustering algorithm, the cluster validity used is Xie-Beni index, introduced by [13] as defined in (10).

$$XB = \frac{\sum_{q=1}^{C}\sum_{r=1}^{N}(\mu_{qr})^2 \left\|v_q - x_r\right\|^2}{N \min_{q\neq r}\left\|v_q - v_r\right\|^2} \tag{10}$$

where N is the number of object, C is the number of cluster, V_q is the q^{th} cluster centre in FCM, μ_{qr} is the membership value and $\|.\|$ is the Euclidean norm. The numerator in (10) is a compactness validity which measures how closely related the objects in the cluster are. A group of measures evaluate cluster compactness based on variance where lower variance indicates better compactness. On the other hand, the denominator in (10) is separation validity which measures how distinct or well separated a cluster is from another cluster. The smaller the separation validity value is, the larger the probabilities that there will be a redundant cluster centre in the clustering.

2.4. Fuzzy linear regression (Tanaka)

The goal of fuzzy linear regression analysis is to determine a regression model which fits all observed fuzzy data within a specified fitting criterion. In identifying the different fuzzy linear regression model, it depends on the fitting criterion used. Based on the Tanaka method, the fuzzy model is a human estimation where it was identified by estimating fuzzy parameters $A_q^* = (\alpha_q, c_q)$ to solve the linear programming problems in (11).

$$\min_{\alpha,c} = c_1 + ... + c_N \tag{11}$$

subject to $c \geq 0$ and

$$\begin{aligned} &\alpha^s x_q + (1-H)\sum_r c_r \left|x_{qr}\right| \geq y_q + (1-H)e_q \\ &-\alpha^s x_q + (1-H)\sum_r c_r \left|x_{qr}\right| \geq -y_q + (1-H)e_q \end{aligned} \tag{12}$$

where α is centre, x is independent variable, H is degree of fitting, c is width, y is dependent variable and e is error. The best fitted model for the given data could be obtained by solving the conventional linear programming problem in (12).

Tanaka's method considered the independent variable value in correspondence to a crisp coefficient does not influence the fuzziness of a predicted value of the dependent variable. It also shows the fuzzy centre α_q and their fuzzy width c_q are scale dependents. In addition, Tanaka's method is sensitive to outliers.

2.5. Fuzzy linear regression (Ni)

Ni approached this model as an extension from fuzzy linear regression model proposed by Tanaka in 1982 [14]. The proposed fuzzy linear regression by Ni are shown in (13) and (14).

$$\min_{\alpha,c} = c_1 + ... + c_N \tag{13}$$

with subject to $c \geq 0$ and

$$\begin{aligned} &\alpha^s x_q + (1-H)\sum_r c_r \left|x_{qr}\right| \geq y_q \\ &-\alpha^s x_q + (1-H)\sum_r c_r \left|x_{qr}\right| \geq -y_q \end{aligned} \tag{14}$$

Where α is the centre and c is the spread of data.

2.6. Fuzzy linear regression (Zolfaghari)

Zolfaghari et al. [6] proposed a new model as an extension of FLR model by Tanaka and Ni, with consideration on fuzzy number and their membership functions. There are two parameter situations which could be used under this model: FLR with symmetric and asymmetric parameter. This study focuses on symmetric parameter which has fuzzy coefficients assumed as TFNs. As indicated, the salient features of the TFNs are its mode, its left and right spread and its support. If the two spreads are equal, the TFN is known as a symmetrical TFN. The proposed FLR by Zolfaghari as shown in (15) and (16).

$$\min = 2ms_0 + 2\sum_{q=1}^{N}[s_q \sum_{r=1}^{m}\left|x_{qr}\right|] \tag{15}$$

With subject to $c \geq 0$ and

$$\begin{aligned} &(1-H)s_0 + (1-H)\sum_{q=1}^{N}(s_q\left|x_{qr}\right|) - a_0 - \sum_{q=1}^{N}(a_q x_{qr}) \geq -y_r \\ &(1-H)s_0 + (1-H)\sum_{q=1}^{N}(s_q\left|x_{qr}\right|) + a_0 - \sum_{q=1}^{N}(a_q x_{qr}) \geq y_r \end{aligned} \tag{16}$$

2.7. A new hybrid of fuzzy c-means method and fuzzy linear regression model

A new hybrid model is defined as a combination of both fuzzy c-means method and fuzzy linear regression model in predicting manufacturing income. Initially, one of the ways to find the best FCM clustering is by choosing the highest correlation value between dependent variable and all independent variables. After that, once a few of higher correlation value can be identified then FCM clustering will be used. Then, the comparison error by MSE, MAE and MAPE will be done and the smallest error value will be chosen.

Next, the best FCM clustering can be identified by calculating *XB*-value that proposed by [13] to reach the minimum value which nearer to zero. Once the optimal cluster number is identified, then proceed to model the data by using FLR (Zolfaghari). This FLR model is used to find the model for each cluster. The best FLR model is identified by getting smallest error value by using three methods of error which are MSE, MAE and MAPE. The degree of fitting is adjusted between 0 and 1 to decide the best model.

3. Results and Discussion

3.1. A new hybrid of fuzzy c-means and fuzzy linear regression

Under fuzzy analysis, there are no assumptions needs to be considered. For the new model, the data were analysed by using MATLAB software to identify the best FCM clustering. The improvisation model is used in this study as it gives better estimation in comparison to other models. This comparison illustrates that the improvisation model achieves the better predicting in manufacturing income. In this study, the dependent variable is the total income (Y) and the independent variables are total salaries and wages paid (x_5), number of degree holder and above (x_6) and total expenditure (x_9). The independent variables x_5, x_6 and x_9 are chosen due to the high correlation value (r) between total income (Y) as compared to the other variables.

The highest correlation value is between x_9 and Y with r = 0.986, followed by the second highest correlation value between x_5 and Y with r = 0.649, and the third highest correlation value between x_6 and Y with r = 0.596. The correlation values and the comparison error values among Y, x_5, x_6 and x_9 could be seen in Table 1 and Table 2 respectively. Table 1 presents the results of the correlation value between the dependent variable and each independent variable. The highest correlation value of x_5, x_6 and x_9 are chosen among the others, as higher correlation value leads to stronger linear correlation. Otherwise, the lowest correlation value leads the weak linear correlation. Therefore, it's not suitable to be used for further analysis of improvisation model.

Table 1: Correlation values of FCM method between Y and x_5, x_6, x_9

	Pearson Correlation (r)
Y vs x_0	0.000
Y vs x_1	0.053
Y vs x_2	0.194
Y vs x_3	0.459
Y vs x_4	0.557
Y vs x_5	0.649
Y vs x_6	0.596
Y vs x_7	0.579
Y vs x_8	0.490
Y vs x_9	0.986

Table 2: Comparison error values of FCM method between Y and x_5, x_6, x_9

	MSE	MAE	MAPE
Y	7.55329000000	176944.0347	119.7584
x_5	7.03966000000	170342.9475	115.0881
x_6	5.73810000000	152709.9697	95.8433
x_9	6.89959000000	168473.8717	113.8235

Table 2 shows the comparison of error values of MSE, MAE and MAPE for variables Y, x_5, x_6 and x_9. It is clearly seen that x_6 was chosen for clustering as it achieved the smallest error values compared to other values with MSE = 5.73810000000, MAE = 152709.9697 and MAPE = 95.8433. The error values was selected among the best correlation value between all variable involved. Therefore, variable x_6 was chosen to determine the best FCM clustering. Table 3 presents the number of *XB*-value for x_6 according to the number of clusters, *c*. The number of clusters chosen are two (c=2) since the *XB*-value reached minimum value nearest to zero of 0.0144 compared to other number of clusters. The *XB*-value of the cluster can be calculated by using Xie-Beni index as in (10).

Table 3: The value of c and XB for x_6

Number of clusters, c	2	3	4	5
XB-value	0.0144	0.0743	0.0532	0.5097

Table 4 shows the details for the variable x_6 according to the cluster. The number of data for cluster 1 is 2827 whereas the number of data for cluster 2 is 29 where for cluster 1, the minimum value for x_6 is 1 whilst the maximum value is 247. Meanwhile, for cluster 2, the minimum value for x_6 is 259 and the maximum value is 1370. FCM model used in this study is to find an improved data which can contribute to a better model with smaller error.

Table 4: Details of the variable x_6

	Cluster 1	Cluster 2
Number of data (N)	2827	29
Minimum value	1	259
Maximum value	247	1370

3.2. Fuzzy linear regression (Zolfaghari) on cluster 1

Once the data is clustered, we then proceed to analyse it by using FLR (Zolfaghari), where the FLR model for cluster 1 is used which involved all the independent variables. This model then evaluated by three measurements of performance which are MSE, MAE and MAPE by adjusted different degree of fitting (H) between 0 and 1 (see Table 5). The smallest error value is considered as the best FLR model for cluster 1 with H = 0.025 as in (17).

$$\hat{Y} = (3500, 0) + (-288, 0)x_1 + (-2212, 0)x_2 + (-0.9130, 0)x_3 + (-44732, 0)x_4 + (1.9245, 0)x_5 + (44241, 0)x_6 + (44534, 0)x_7 + (44729, 0)x_8 + (2.4669, 2.1163)x_9 \quad (17)$$

The fuzzy parameter results as in Table 6 indicated that the imprecision of manufacturing income can be represented by the parameter (x_9) which is the total expenditure of 2.1163. The error values of FLR (cluster 1) for MSE, MAE and MAPE are shown in Table 5 with 182419000000, 114508.0207 and 95.8043 respectively.

Table 5: Measurement error for FLR model (Zolfaghari-cluster 1)

H	MSE	MAE	MAPE
0.025	182419000000	114508.0207	95.8043
0.05	190152000000	121726.6229	110.2542
0.1	191827000000	122315.4418	110.9928
0.15	193424000000	122737.5068	111.1340
0.185	194348000000	122790.9518	110.6683
0.2	195093000000	123283.6196	111.6638
0.3	198360000000	124201.6926	112.1815
0.4	201724000000	125232.4220	113.0034
0.5	205260000000	126564.5333	115.0010
0.6	208568000000	127364.9507	115.0683
0.7	211945000000	128255.6598	115.4368

0.8	215400000000	129260.8907	116.2600
0.9	218948000000	130161.7870	116.7546

Table 6: Fuzzy parameter for cluster 1 (*H*=0.025)

Fuzzy Parameter	Fuzzy Center, α_q	Fuzzy Width, c_q
x_0	3500	0.0000
x_1	-288	0.0000
x_2	-2212	0.0000
x_3	-0.9130	0.0000
x_4	-44732	0.0000
x_5	1.9245	0.0000
x_6	44241	0.0000
x_7	44534	0.0000
x_8	44729	0.0000
x_9	2.4669	2.1163

3.3. Fuzzy linear regression (Zolfaghari) on cluster 2

The FLR model for cluster 2 is modelled involving all independent variables. This model is evaluated by three measurements of performance with different degree of fitting (*H*) is adjusted between 0 and 1 (see Table 7). The smallest error value reveal that as the best model for cluster 2 with *H* = 0.05 as in (18).

$$\hat{Y} = (-99999, 0) + (-99999, 0)x_1 + (26531, 0)x_2 + (-0.1484, 0)x_3$$
$$+(-46499, 0)x_4 + (5.9468, 0)x_5 + (46605, 0)x_6 + (46560, 0)x_7$$
$$+(46454, 0)x_8 + (0.9905, 0.0246)x_9$$

The results of fuzzy parameters are given in Table 8, indicated that the imprecision of manufacturing income can be represented by the fuzziness of parameter (x_9) which is the total expenditure of 0.0246. The smallest error values of FLR (cluster 2) are MSE=190057000000, MAE=254814.5620 and MAPE=20.1972.

Table 7: Measurement error for FLR model (Zolfaghari-cluster 2)

H	MSE	MAE	MAPE
0.025	1.90237E+11	255074.3838	20.1905
0.05	1.90057E+11	254814.5620	20.1972
0.1	1.90140E+11	255022.4454	20.3557
0.15	1.90123E+11	255137.3197	20.2824
0.185	1.90068E+11	254951.1805	20.2839
0.2	1.90031E+11	254815.7883	20.2832
0.3	1.90101E+11	255098.8480	20.3736
0.4	1.90051E+11	256155.1067	20.4897
0.5	9.12315E+16	279409164.4	19870.2031
0.6	3.63977E+17	405973147.0	16110.0865
0.7	1.78239E+31	3.92151E+15	2.4423E+11
0.8	4.21207E+16	179421578.2	12908.5701
0.9	1.84994E+11	275862.0251	21.7661

Table 8: Fuzzy parameter for cluster 2 (*H*=0.05)

Fuzzy Parameter	Fuzzy Center, α_q	Fuzzy Width, c_q
x_0	-99999	0.0000
x_1	-99999	0.0000
x_2	26531	0.0000
x_3	-0.1484	0.0000
x_4	-46499	0.0000
x_5	5.9468	0.0000
x_6	46605	0.0000
x_7	46560	0.0000
x_8	46454	0.0000
x_9	0.9905	0.0246

Table 9: Summary of measurements of error of models (18)

	MSE	MAE	MAPE
MLR	5677188424000	199997.4805	135.9859
FLR (Tanaka, H=0.95)	829647000000	185539.8661	125.5615
FLR (Ni, H=0.10)	815481000000	183730.5705	124.0696
FLR (Zolfaghari, H=0.025)	575619000000	154163.8335	105.6919
Improvisation – Cluster 1 (H=0.025)	182419000000	114508.0207	95.8043
Cluster 2 (H=0.05)	190057000000	254814.5620	20.1972
Total overall error value improvisation model	182496556700	115932.702	95.0366

Table 9 presents the summary of error values for all model involved in this study which is MLR model, FLR model proposed by (Tanaka, Ni and Zolfahari) and a new hybrid model of FCM and FLR. Each value of error was measured by using MSE, MAE and MAPE. It's clearly shown that a new hybrid model reached the smallest error value for cluster 1 (*H*=0.025) and cluster 2 (*H*=0.05) compared to other models. The value of MSE, MAE and MAPE for cluster 1 is 182419000000, 114508.0207 and 95.8043 respectively whereas, for cluster 2 is 190057000000, 254814.5620 and 20.1972 respectively. Meanwhile, the total overall error of MSE, MAE and MAPE for a new hybrid model is 182496556700, 115932.702 and 95.0366 respectively.

4. Conclusion

This research is executed to improvise a certain model which could be used in predicting manufacturing income for industry company. The first objective was achieved through fuzzy c-means (FCM) method. A few variables with higher correlation value was selected to further analysed using FCM clustering toward manufacturing income data. After that, the MSE, MAE and MAPE will be compared among related variables. Among the all variables involved, x_6 was chosen according to its smallest error value. In analysing this FCM clustering, Xie-Beni index is used as a validity index. It was used to determine how to better data was obtained in various clusters. Cluster number of two (*c*=2) was chosen since it has the lowest *XB*-value, which is 0.0144. The previous researcher method by [15] indicated that the traditional (K-means)

algorithm seem to be superior in cluster the data. However, it may not be able to find overlapping cluster for large datasets. Therefore, FCM clustering should be provided because it can handle the large datasets and able to allow an item to belong to more than one cluster.

The second objective was achieved by means of existing methods which are MLR model and fuzzy linear regression model that proposed by Tanaka, Ni and Zolfaghari. The results of each model error value were discussing in the previous chapter on data exploration section to see the comparing value. The model with the smallest error value will be used in the next improvisation model. Based on results obtained, fuzzy linear regression model (Zolfaghari) reached smaller error value when compared with another model. The fuzzy regression analysis might be very widely applied in various datasets but according to each author it still has differences among them in terms of linear programming. Therefore, there are difference results under fuzzy regression analysis. In this study, model proposed by [6] indicated the smallest error value for MSE, MAE and MAPE with adjusted the degree of fitting (*H*). This decrease in the modelling error brings forward the opportunity in predicting manufacturing income of industry company with more effectively.

The third objective was achieved by a new hybrid model of fuzzy c-means method and fuzzy linear regression model towards predicting manufacturing income data. Result from the first objective have been presented the chosen number of clusters of c=2 by using FCM method with cluster 1 and 2 having an amount of 2827 and 29 data respectively. Meanwhile, the result of second objective for modelling was chosen fuzzy linear regression (Zolfaghari) before using the improvisation model. The finalise result of both cluster for MSE, MAE and MAPE value has combined by calculated as a total overall error value which are 182419000000, 115932.702 and 95.0366 respectively. In selecting the smallest error value, it will increase the profitability and may reduce losses for an industry company. At the same time, industry planning will come more accurate and effective to predict exact income. So, this improvise model is suitable to develop a potential model in making next prediction. The result from improvisation model of FCM method and FLR model can predict the manufacturing income with input values from all independent variables is available for the coming prediction. Therefore, it can be concluded that the improvisation model of FCM method and FLR model is the best method in predicting manufacturing income.

To characterise the best model traits in statistical method analysis, the fourth objective could be achieved by using three measurements of error including mean square error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). These three measurements of error were measured against each model involved in this study to discover the model having smallest error value. The smaller error value indicates more accurate and stable prediction than other [16]. Measurements of error for each model were evaluated by optimising different degree of fitting (*H*). The degree of fitting of measurements of error could be optimised between 0 and 1. Once the smallest error value is identified, the degree of fitting (*H*) is optimised to find the best model. As has been proved, the improvisation model achieves the smallest error value for the three measurements error among the other model involved in this study. Based on the results, a new hybrid model can be employed to produce manufacturing income predictions that would be extremely useful for any industry company. This model also be promising model that offers high precision toward predict exact income.

Acknowledgement

The authors would like to express gratefully heartfelt thanks to the Universiti Tun Hussein Onn Malaysia and Office for Research, Innovation, Commercialization and Consultancy Management (ORICC) for the financial support under the TIER 1 research grant (Vot H232).

References

[1] Kaufman L & Rousseeuw PJ, Finding Groups in Data: An Introduction to Cluster Analysis, Canada: A Wiley-Interscience Publication (1989).

[2] Everitt BS, Landau S, Leese M & Stahl D, Cluster Analysis. 5th ed., King's College London: A John Wiley and Sons Ltd. Publication (2011).

[3] Aggarwal CC & Reddy CK, Data Clustering: Algorithms and Applications, London: CRC Press Taylor & Francis Group (2014).

[4] Chaudhuri A & De K, Achieving Greater Explanation Power and Forecasting Accuracy with Non-uniform spread Fuzzy Linear regression, School of Science, Netaji Subhas Open University, India (2013).

[5] Tanaka H, Uejima S & Asai K, "Linear Regression analysis with fuzzy model", IEEE Transactions on Systems Man and Cybernetics, 12, (1982) 903-907.

[6] Zolfaghari ZS, Mohebbi M & Najariyan M, "Application of fuzzy linear regression method for sensory evaluation of fried donut. Ferdowsi University of Mashhad, Iran", Applied Soft Computing 22, (2014) 417-423.

[7] Mendenhall W, Beaver RJ & Beaver BM, Introduction to Probability and Statistics (14th ed.), Pacific Grove, Califonia (2013).

[8] Garcia C, Gomez RS, Perez JG & Martin, MDML, "On the selection of the ridge and raise factors", Indian Journal of Science and Technology, 10(13) (2017).

[9] Dunn J, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters", Journal of Cybernetics, 3, (1973) pp. 32-57.

[10] Bezdek JC, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York, (1981) doi: 10. 1007/978-1-4757-0450-1.

[11] Maulik U & Bandyopadhyay S, "Performance evaluation of some clustering algorithms and validity indices", IEEE PAMI, 24, (2002) pp. 1650-1654.

[12] Tan PN, Steinbach M & Kumar V, Introduction to Data Mining. USA: Addison-Wesley Longman, Incorporated (2005).

[13] Xie XL & Beni GA, "Validity Measure for Fuzzy Clustering", IEEE Transactionson Pattern Analysis and Machine Intelligence, 13(4), (1991) pp. 841-846.

[14] Ni Y, Fuzzy Correlation and Regression Analysis. University of Oklahoma Graduate College, UMI number : 3163014 (2005).

[15] Ghosh S & Dubey SK, "Comparative Analysis of K-mean and Fuzzy C-means Algorithms", (IJACSA) International Journal of Advanced Computer Science and Applications, 4(4) (2013).

[16] Xiao L, Shao W, Liang TL & Wang, CA. "Combined model based on multiple seasonal patterns and modified firefly algorithm for electrical load forecasting", Application Energy, 167:135e53 (2016).

[17] Samat NA, Salleh MNM, "A study of data imputation using fuzzy c-means with particle swarm optimization", Advances in Intelligent Systems and Computing, 549 AISC (2017), 91-100

[18] Hussain K, Mohd Salleh MN, "Analysis of techniques for ANFIS rule-base minimization and accuracy maximization", ARPN Journal of Engineering and Applied Sciences, 10 (20), (2015), 9739-9746.

[19] Salleh MNM, "A fuzzy modelling of decision support system for crop selection", ISIEA 2012 - 2012 IEEE Symposium on Industrial Electronics and Applications, 6496622 (2012), 17-22.

[20] Khalid K, Mohamed I and Abdullah NA, An Additive Outlier Detection Procedure in Random Coefficient Autoregressive Models AIP Conference Proceedings 1682, (2015), 050017.

[21] Mohamed I, Khalid K And Yahya MS, Combined Estimating Function for Random Coefficient Models with Correlated Errors Communications In Statistics—Theory And Methods 45(4), (2016), 967-975.