Fuzzy model for clustering open pollinated maize variety released in Indonesia

Muhammad Aqil, N.N. Andayani, Fahdiana T and Suwardi
Indonesian Cereals Research Institute, Indonesia
Email: aqil@pertanian.go.id

Abstract. Open-pollinated variety (OPV) is a potential gene source for developing new high yielding maize hybrid. Grouping of maize OPV varieties requires a model analysis tool such as the fuzzy clustering in order to identify the similarity measure. The objective of the research was to characterize the similarity among OPV varieties for further research including site-specific OPV development. As many as 35 OPV varieties were assessed and clustered into groups by using fuzzy c means algorithm. The result indicated that fuzzy model clustering enabled to develop a relationship among the OPV maize and morphological traits. Based on the similarity among agronomic characters, OPV maize varieties can be clustered into four group. Group I represent high productivity, drought and disease tolerant varieties. Group II contains maize varieties characterized by late silking and maturity days. Further, Group III represent varieties with the earliest maturity varieties and low grain yield. Group IV indicates the OPV maize having moderate grain yield and susceptible to downy mildew disease. The information on similarity among variety or group can be used for varietal improvement programs such as recombination and molecular based gene introgression not only for maize but also other crops. Our findings also provide insights to understand similarity among each variety, which are considered to be valuable for future OPV breeding programs such as drought, lodging and disease resistant varieties.

1. Introduction
Population improvement through the development of lines, OPVs, and hybrids is a continuous process and well organized in Indonesia. Since future maize expansion will very much rely on the dryland ecologies of outer islands, site specific oriented breeding programs for low-productivity or stressed environments should be initiated. Maize improvement and agronomy research in Indonesia is now mainly carried out by Indonesian Agency for Agricultural Research and Development (IAARD). From this institute, it has been released 46 open pollinated as well as over 40 hybrid maize [1]. Each variety has its own advantages and similarity that can be utilized for breeding or dissemination purposes.

In general, the grouping of maize OPV varieties requires a model analysis tool such as the fuzzy clustering and prediction techniques. Clustering essentially deals with the task of splitting a set of patterns into a number of more or less homogenous classes (or clusters) with respect to a suitable similarity measure such that the patterns belonging to any one of the clusters are similar and the patterns of different clusters are as dissimilar as possible [2]. There are many clustering methods available, and each of them may give a different grouping of a dataset, such as (hard) c-means (or k-means) clustering, fuzzy c-means clustering, the mountain clustering, subtractive clustering, etc. Fuzzy c-means clustering is a data clustering algorithm in which each data point belongs to a cluster to a degree specified by a membership grade [3]. The fuzzy c-means algorithm has proved to be effective in the exploratory
detection of data structures, and its popularity has grown considerably among researchers in diverse fields.

Successful applications of this method in detection and classification of object patterns have been widely reported. [4] used a rule base system using fuzzy logic for generating a prediction of land used for farming, rainfall and production. [5] analyzed pest and disease classifications for paddy based on Fuzzy k-means (Fkm) and Fuzzy c-means (Fcmm). They found that fuzzy c-means clustering analysis produced clear membership patterns and classification among paddy pests and diseases effectively. A modeling approach by using Fcm was also introduced by [6] to help farmers making decisions in precision agriculture. [7] and [8] demonstrated the use of fuzzy c-means to discriminate the crops (paddy, maize, sugarcane, and cotton). [9] used segment incomplete nutrient-deficient crop fuzzy c-means based on color clustering. Finally, [10] investigated plant phenotyping (stereo vision) by using three dimension point cloud segmented by spectral clustering. [11] used fuzzy algorithm for clustering land used for maize, soybean and rice in Southeast Minahasa Indonesia.

This paper describes an attempt to examine the potential of the fuzzy c-means algorithm for characterization the similarity among OPV varieties for further research including site specific OPV and hybrid breeding programs.

2. Material and Methods

2.1. Maize OPV Dataset

The original data set of the OPV maize variety consists of yield and yield attributed traits were collected from three institutes viz., Center for Plant Variety Protection and Agriculture Licensing, Indonesian Cereals Research Institute, and Indonesian Agency for Research and Development. The dataset was obtained by taking the last 56 years of released OPV in Indonesia (1956-2011). Genetic source, year released and specific advantage of each variety of OPV maize released in Indonesia is shown in Table 1.

In this study, OPV maize varieties having similar agronomic performance would be expected to behave similarly. The entire data set containing larger number of traits including grain yield, plant height, ear height, ear aspect, plant aspect, silking days, anthesis to silking days, maturity, 1000 kernel weight and disease resistance were first analyzed. Furthermore, the larger dataset will be reduced to a selected fewer number of traits of varieties (viz., grain yield, silking days, maturity, 1000 kernel weight and disease resistance) which perform significantly different response pattern and distinguishing the variation among traits analyzed. Reducing the number of traits used in analysis enabling easier interpretation of clustered traits. As for downy mildew, a quantitative approach was adopted to convert the qualitative values (downy mildew resistance) into quantitative values.

| Variety          | Genetic source         | Year released | Advantage               |
|------------------|------------------------|---------------|-------------------------|
| Metro            | Tequisate Guatemala    | 1956          | Suitable for <1,000 m asl|
| Baster Kuning    | Margaayu-Priangan      | 1950-1960     | Suitable for highland   |
| Kania            | Kenya Afrika           | 1950-1960     | Suitable for food       |
| Malin            | Introduced from British| 1950-1960     | Early maturity          |
| Harapan          | Guatemala              | 1964          | Suitable for highland   |
| Bima             | South America          | 1966          | Suitable for highland   |
| Pandu            | Kenya Africa           | 1966          | Suitable for food       |
| Permadi          | Philippines-Guatemala  | 1966          | Suitable for lowland    |
| Bogor Komposit 2 | Introduced from India  | 1969          | Good husk cover         |
| Harapan Baru     | Phil-DMRS              | 1978          | Disease tolerant        |
| Arjuna           | Introduced from Thailand| 1980        | Suitable for highland   |
| Bromo            | Philippines            | 1980          | Suitable for food       |
2.2. Similarity Analysis

The main objective of using fuzzy clustering in the analysis of data from maize varieties is to group the varieties into several homogeneous groups such that those varieties within a group have a similar pattern. The fuzzy function will map the design parameter to membership grade between that of scaled interval (Figure 1).

Fuzzy c-means clustering will locate each data point belongs to a cluster with a degree specified by membership grade [12]. This method is an unsupervised classification algorithm for iteratively determining the local minima in the objective function [13]. The objective function, which is minimized iteratively, is a weighted within-groups sum of distances $d_{i,k}$. The weighting has been done by multiplying the squared distances by membership values $u_{i,k}$. An important parameter in this algorithm is the exponent $m$. The total number of clusters is given by $c$ while for object in calibration data is given by $n$. Fuzzy c-means is an iterative algorithm for minimizing the objective function.

$$J_m(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{i,k})^m (d_{i,k})^2$$
where $U$ is the membership matrix, and $V$ is the cluster centers matrix. The $i$th cluster center is symbolized by $v_i$ and the vector of measurements for the data point $k$ by $x_k$, then Euclidean distance was adopted for calculating distance of cluster. Cluster centers are calculated as follow:

$$
v_i = \frac{\sum_{k=1}^{n} (u_{i,k})^m x_k}{\sum_{k=1}^{n} (u_{i,k})^m}, \forall i
$$

Convergence can be calculated by comparing the changes in the membership function or the cluster center at two successive iterations. All data collected for yield and yield related component parameters were analyzed using MATLAB Software [14].

3. Results and Discussion

Prior to grouping, fuzzy will first optimize the model parameters such as grain yield, silking days, maturity, 1000 kernel weight and disease resistance. Optimization is iteratively classifying each maize variety and moves the varieties among groups until it finds the most satisfying group. The cluster centers were calculated in an unsupervised way, where the fuzzy partitioning provided allows a given data point belong to one or more cluster partially (Fig. 2). Moreover the fuzzy c-means algorithm generated cluster descriptions and the degrees of membership (varies between 0 and 1) for belonging to the different clusters.

In order to initialize minimization, the values of the exponent ($m$) and the number of clusters ($c$) were fixed at 2 and 4 respectively. In addition, the mean squares error (MSE) was used as our stopping criteria with a value of $1 \times 10^{-5}$. After that, the algorithm was allowed to find the optimal partition, i.e. to minimize the weighted sum of distances within groups. In other words, the objective function is minimized when high membership values are obtained in areas where the observations are close to the centroid, and low membership values are obtained where observations are distant from the centroid. The clustering algorithm stops as soon as the absolute value of differences of all pairs of elements in a successive pair of $U$ matrices differs by less than $1 \times 10^{-5}$. From the algorithm runs, the optimal solution was found after approximately 55 iteration counts with minimized objective function value of 25600.30 (Figure 2). The number of iteration can be longer depend on the output that will be executed [15].

![Figure 2. Iteration process to fit the objective function (left) and the result of classification (right)](image-url)
Table 2. Fuzzy membership matrix for all data samples

| Variety                | Membership value/Cluster membership |
|------------------------|-------------------------------------|
|                        | I        | II       | III      | IV       |
| Metro                  | 0.008    | (0.580)  | 0.329    | 0.084    |
| Baster Kuning          | 0.141    | (0.752)  | 0.047    | 0.061    |
| Kania                  | 0.017    | (0.944)  | 0.012    | 0.028    |
| Malin                  | 0.221    | (0.743)  | 0.004    | 0.033    |
| Harapan                | 0.086    | (0.853)  | 0.034    | 0.028    |
| Bima                   | 0.167    | 0.220    | 0.089    | (0.523)  |
| Pandu                  | 0.021    | (0.923)  | 0.014    | 0.043    |
| Permadi                | 0.001    | (0.982)  | 0.009    | 0.009    |
| Bogor Komposit 2       | 0.040    | 0.003    | 0.010    | (0.946)  |
| Harapan Baru           | (0.893)  | (0.635)  | 0.298    | 0.061    |
| Arjuna                 | 0.006    | 0.003    | (0.750)  | 0.030    |
| Bromo                  | 0.008    | 0.002    | (0.990)  | 0.002    |
| Abimayu                | 0.175    | 0.015    | (0.745)  | 0.066    |
| Manding                | 0.007    | 0.003    | (0.760)  | 0.230    |
| Talango                | 0.006    | 0.004    | (0.790)  | 0.200    |
| Nakula                 | 0.007    | (0.879)  | 0.110    | 0.004    |
| Parikesit              | 0.006    | (0.785)  | 0.066    | 0.143    |
| Sadewa                 | 0.004    | (0.873)  | 0.060    | 0.063    |
| Piet kuning            | 0.009    | (0.732)  | 0.109    | 0.151    |
| Kalingga               | 0.385    | 0.011    | 0.062    | (0.542)  |
| Wiyasa                 | (0.848)  | 0.004    | 0.082    | 0.066    |
| Antasena               | (0.969)  | 0.008    | 0.018    | 0.011    |
| Wisanggeni             | (0.880)  | 0.004    | 0.051    | 0.066    |
| Bisma                  | 0.226    | 0.010    | 0.046    | (0.718)  |
| Surya                  | 0.278    | 0.010    | 0.051    | (0.661)  |
| Lagaligo               | (0.879)  | 0.004    | 0.051    | 0.066    |
| Gumarang               | 0.140    | 0.005    | (0.797)  | 0.058    |
| Lamuru                 | (0.981)  | 0.005    | 0.086    | 0.007    |
| Kresna                 | (0.883)  | 0.002    | 0.088    | 0.028    |
| Bayu                   | (0.877)  | 0.003    | 0.008    | 0.032    |
| Palakka                | (0.986)  | 0.000    | 0.119    | 0.005    |
| Sukmaraga              | (0.813)  | 0.005    | 0.082    | 0.062    |
| Srikandi Kuning        | (0.848)  | 0.004    | 0.010    | (0.950)  |
| Provit A1              | 0.037    | 0.003    | 0.010    | (0.950)  |
| Anoman                 | 0.070    | 0.030    | 0.050    | (0.920)  |

The portion of the fuzzy c-means membership matrix for all OPV varieties are shown in Table 2. Results shown in Table 2 were then used to map the maize OPV variety according to the degree of similarity. As an illustration, the membership values of Metro variety in Table 2 were found to be 0.008 (Group I), 0.580 (Group II), 0.329 (Group III) and 0.084 respectively. From those four values, therefore Metro variety was more likely to belong to Cluster II (or Group II) since it has the highest degree of membership to this cluster compared to the other three. In addition to Provit A1 (high beta carotene maize) variety, calculation of membership values were: 0.037 (Group I), 0.003 (Group II), 0.010 (Group III) and 0.950 (Group IV), thus Provit A1 maize may belong to Cluster IV/Group IV. By the same interpretation, the widespread OPV variety according to their clusters can be classified, and the results
obtained is presented in Table 3. The varieties belong to the same group have similar agronomic characteristics while different group indicate dissimilar character.

Table 3. Classification of maize variety based on the similarity characteristic

| Group II          | Group I                  |
|-------------------|--------------------------|
| Metro             | Srikandi Kuning          |
| Permadi           | Sukmaraga                |
| Kania             | Palakka                  |
| Pandu             | Bayu                     |
| Parikesit         | Kresna                   |
| Sadewa            | Lamuru                   |
| Harapan           | Lagaligo                 |
| Piet Kuning       | WIsanggeni               |
| Baster Kuning     | Antasena                 |
| Malin             | Wiyasa                   |
|                   |                          |
| Group III         | Group IV                 |
| Gumarang          | Bisma                    |
| Bromo             | Provit A                 |
| Arjuna            | Kalingga                 |
| Abimanyu          | Bima                     |
| Nakula            | Surya                    |
| Manding           | Bogor Komposit II        |
| Talango           | Harapan Baru             |
|                   | Anoman                   |

The matrix of cluster centers enabled us to develop a relationship among the OPV maize and morphological traits. Table 4 shows the representative parameter values of the four cluster centers (or four OPV maize groups) derived from fuzzy model. Once the cluster centers have been estimated, the classification of the level of each term of maize variety features among those four groups (or clusters) can be determined [16]. The classification results indicate that there is a clear sequence in the four calculated cluster centers. As shown in Fig. 3 and Table 3, the clustering results comprised four groups: one group with low yield and latest maturity, one group with high yield and disease tolerance, while the remained two groups were less moderate and moderate yield and agronomic components. The brief explanation of the characteristics of each group is given as follows.

Table 4. Cluster centers for fuzzy clustering of OPV variety

| Term                     | Cluster center values |
|--------------------------|-----------------------|
|                          | I         | II        | III       | IV        |
| Silking (50%) (days)     | 54.511    | 80.083    | 51.271    | 59.916    |
| Maturity (days)          | 95.435    | 140.173   | 89.921    | 107.411   |
| 1000 kernel weight (g)   | 276.566   | 445.356   | 239.095   | 319.993   |
| Disease resistance       | 2.179     | 2.498     | 2.278     | 2.660     |
| Grain yield (t/ha)       | 4.745     | 3.501     | 3.508     | 4.374     |
Group I represents Srikandi Kuning, Sukmaraga, Palakka, Bayu, Kresna, Lamuru, Lagaligo, Wisanggeni, Antasena and Wiyasa variety. The group contain maize varieties with specific advantages such as high grain yield and drought tolerant (Lamuru variety). High yield and early maturity are also among the most important characteristics for choosing a maize variety in Indonesia [17], [18]. Such early types of varieties are appropriate in area with short rainy season so as to escape moisture depletion encountered at stages of crop growth particularly grain filling period or late in the season. Some varieties such as Lagaligo and Wisanggeni also shows a good resistance to downy mildew disease [19]. Group II consists of Metro, Permadi, Kania, Pandu, Malin, Baster Kuning, Parikesit, Sadewa, Harapan and Piet Kuning variety. The group represents mainly maize varieties released between 1950-1980 and characterized by lowest productivity, below 3.5 t/ha. However, this group represents the highest 1000 kernel weight. Kernel weight is particularly important for calculating seed size variations and estimating shattering loss. The final weight achieved by maize kernels is largely genetically determined [20]. Kernel weight has been shown to vary with kernel number per plant. [21] reported that crossing of maize inbreds based on 1000-grain weight can give efficient results. The Group II also contain maize varieties with late silking and maturing days (> 80 days after planting). [22] reported that both dominance and epistatic interaction played a major role in governing inheritance of pollination, silking and maturity.

Further, Group III in which represents Bromo, Abimanyu, Nakula, Manding, Talango variety perform specific properties such as lower grain yield, lowest 1000 kernel weight, earliest silking days and maturity as well as good tolerance to downy mildew, early maturity. Some of those varieties were planted as bird feedplant. These varieties are suitable for planting under short and erratic rainfall condition. Having maturity < 89 days, these varieties can be used as a parental material for breeding of early maturity maize. Another characteristic of this group is the lowest 1000 kernel weight. The two main factors thought to control kernel sink capacity in maize kernel are endosperm cell number and starch granule number [23]. [24] indicated that final seed size also depends on cell expansion. Group IV indicates the list of OPV maize having moderate grain yield, susceptible to downy mildew, slightly high 1000 kernel weight, late silking days and late maturity. The results of clustering indicated that Bisma, Provit A, Kalingga, Bima, Surya, Bogor Komposit II, Anoman and Harapan Baru are comparatively late maturity and some of the members of this group such as Anoman variety is susceptible to downy mildew. However, Bisma variety that is classified in this group is the most widely planted OPV in Indonesia during 2010-2015 [25]. This group also produce a high fresh biomass [26]. This study provides insights to understand similarity among each variety, which are considered to be valuable for future OPV breeding programs in Indonesia. As an example, Gumarang variety (Group III/early maturity maize group) could be selected as a potential line for breeding early maturity maize in Indonesia.

4. Conclusion
The present study provided insights into similarity among maize OPV released in Indonesia from 1950 until 2011. As many as 35 OPV varieties were assessed and clustered into groups by using fuzzy c means algorithm. The result indicated that based on the similarity among agronomic characters, all varieties can be clustered into four group. Group I which contain 10 OPV varieties represents high productivity, drought and disease tolerant maize OPV. Group II contained 10 maize varieties with late silking and maturing days. Further, Group III which contain seven OPV varieties represents varieties with the earliest maturity, relatively low grain yield, lowest 1000 kernel weight and good tolerance to downy mildew. Group IV (eight OPV maize varieties) indicates the OPV maize having moderate grain yield and susceptible to downy mildew disease. The information on similarity among variety or group can be used for varietal improvement programs such as recombination and molecular based gene introgression.
Acknowledgement
The authors would like to thank the Seed Resources Management Unit (UPBS) and Indonesian Cereals Research Institute for providing financial and data support.

References
[1] IAARD.(2017) Description of high yielding maize varieties in Indonesia. Indonesian Agency for Research and Development, Jakarta.
[2] Irani J., Pise N. & Phatak IM. (2016) Clustering Techniques and the Similarity Measures used in Clustering. International Journal of Computer Applications Volume 134 – No.7, January 2016 9.
[3] Ghosh S. & Dubey SK.(2013) Comparative Analysis of K-Means and Fuzzy C-Means Algorithms, IAACSA, Vol. 4, No.4, 2013, pp. 35-39.
[4] Kumar H. & Kumar S. (2015) Prediction of Production of Crops using K-mean & Fuzzy Logic. Journal of Computer Science and Information Technology ISSN 2320–088, Vol. 4, Issue. 8 pp.44-56.
[5] Papageorgiou E.I. (2009) A novel approach on designing augmented Fuzzy Cognitive Maps using fuzzified decision trees”, In: Proceedings at 4th international Symposium of Intelligence Computations and Applications, ISICA.
[6] Palamuleni, L., Annegarn, H., Kneen, M., & Landmann, T. (2007) Mapping rural savanna woodlands in Malawi: a comparison of maximum likelihood and fuzzy classifiers. In Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International pp. 1260-1264). IEEE.
[7] Musande, V., Kumar, A., Roy, P. S. & Kale, K. (2013) Evaluation of fuzzy-based classifiers for cotton crop identification. Geocarto International, 28(3), 243-257.
[8] Balasubramaniam, P. & Ananthi, V.P. (2015) Segmentation of nutrient deficiency in incomplete crop images using intuitionistic fuzzy c-means clustering algorithm. Nonlinear Dynamics 83:849–866
[9] Santos, T.T., Koenigkan, L.V., Barbedo, J.G.A., Rodrigues, G.C. (2014) 3D Plant Modeling: Localization, Mapping and Segmentation for Plant Phenotyping Using a Single Hand-held Camera. In: Computer Vision - ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12: 247–263
[10] Tamaela, E, Sediyono & A. Setiawan (2017). Cluster Analysis Menggunakan Algoritma Fuzzy C-means dan K-meansUntuk Klusterisasi dan Pemetaan Lahan Pertanian. Jurnal Buana Informatika, Volume 8, Nomor 3, Juli 2017: 151-160
[11] Jang, J.S.R., Sun, C.T., & Mizutani, E. (1997) Neuro fuzzy and soft computing: A computational approach to learning and machine intelligence, Prentice Hall, New Jersey, pp.503-534.
[12] Bezdek, J.C. (1980) A convergence theorem forthe fuzzy ISODATA clustering algorithms, IEEE Transaction -RAMI; 2(1).pp.1-8.
[13] Agbonifo O.C & Olufolaji D.B. (2012) A fuzzy expert system for diagnosis and treatment of maize plant diseases. International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, No. 12, pp. 83-89, 2012.
[14] Casa R, Baret F, Buis S, Lopez-Lozano R, Pascucci S, Palombo A. & Jones HG.(2010) Estimation of maize canopy properties from remote sensing by inversion of 1-D and 4-D models. Precision Agriculture. Available from doi:10.1007/s11119-010-9162-9.
[15] Sa M.K., Hong T.K. & Lee J.K. (2018) Genetic Diversity and Association Analyses of Canadian Maize Inbred Lines with Agronomic Traits and Simple Sequence Repeat Markers. Plant Breeding and Biotechnology. 6 (2):159-169.
[16] Pswarayi A & Vivek BS. (2007) Combining ability amongst CIMMYT’s early maturing maize (Zea mays L.) germplasm under stress and non-stress conditions and identification of testers. Euphytica 162:353-362.
[17] Azrai M, Mejaya M.J.& Aswidinnoor H. (2013) Combining ability of QPM maize. Journal of Food Crops Research 33(3): 137-147.
[18] Aqil, M. & Rahmi YA. (2016) Description of high yielding maize varieties in Indonesia. Indonesian Cereals Research Institute, Maros. 2016.
[19] Reddy V.M. & Daynard T.B. (1983) Endosperm characteristics associated with rate of grain filling and kernel size in corn. Maydica 28: 339–355.
[20] Altinbas, M. & M. Tosun. (1998) A study on utilizing combining ability covariances in hybrid maize breeding. Anadolu, 8 (2):90-100.
[21] Sher H., Iqbal M., Khan K., Yasir M. & Rahman HU. (2012) Genetic analysis of maturity and flowering characteristics in maize (Zea maysL.). Asian Pacific Journal of Tropical Biomedic. 2 (8): 621–626.
[22] Jones, R. J., B. M. N. Schreiber, & J. A. Roessler. (1996) Kernel sink capacity in maize: Genotypic and maternal regulation. Crop Science 36: 301–306.
[23] Egli & Tekrony, D.M. (1997) Species differences in seed water status during seed maturation and germination. Seed Science Research 7: 3–11.
[24] Directorate General of Seed Service.(2016) Inventarization of variety spread across Indonesia. Annual Report, Jakarta 2016.
[25] Suwardi & Roy Efendi. (2009) Efisiensi penggunaan pupuk N pada jagung komposit menggunakan bagan warna daun. Prosiding Seminar Nasional Serealia.