Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 10, No. 2
journal.magisz.org

Comparison of Internal Validity Indices for Fuzzy Clustering

Zeynel Cebeci[1]

INFO
Received 30-11-2019
Accepted 16-12-2020
Available online 28-01-2020
Responsible Editor: M.Herdon:

Keywords:
cluster analysis, fuzzy clustering, number of clusters, cluster validation, internal validity index.

ABSTRACT

Partitioning clustering has been one of the key components of data analytics to discover meaningful patterns in agricultural big data, driven by the increasing use of IoT-based technologies in smart farming. In partitioning clustering, the quality of clustering or performances of clustering algorithms are mostly evaluated by using the internal validity indices. In this study, the effectiveness of some widely used internal fuzzy indices are compared using the basic Fuzzy C-Means clustering algorithm. It is especially aimed to investigate changes in the effectiveness of validity indices when fuzzy data points are at different distances from the cluster centers. According to the results obtained on the simulated two-dimensional datasets, Fuzzy Silhouette, Fuzzy Hypervolume and Kwon are the most successful indices in validation of fuzzy clustering results.

1. Introduction

In data mining for knowledge discovery, clustering is one of the most widely used unsupervised learning techniques to explore the meaningful substructures or patterns in examined datasets. Clustering as an exploratory data analysis is frequently applied in almost every area of agriculture, food, environment and the other life sciences subjects ranging from genomics to biomedical image segmentation. In recent years, clustering has gained an increasing importance in knowledge discovery from the big data collected via agricultural data acquisition systems and sensors networks based on the systems using IoT with special reference to precision agriculture (Vendrusculo & Kaleita 2011; Cao *et al* 2012; Tian & Li 2015; Bangui *et al* 2018, Marcu *et al* 2019). Recently, Majumdar *et al* (2017) analyzed the agricultural big data for finding optimal parameters to maximize the crop production using clustering based data mining techniques. Since the agricultural data frequently exhibit fuzzy characteristics, the use of the fuzzy algorithms and validation techniques are required for clustering applications on agricultural datasets.

Cluster analysis aims to divide a dataset into *c* (or *k*), numbers of groups by using the hierarchical or non-hierarchical clustering algorithms. As a result of clustering, similar set of data points are brought together to form subgroups which are called as clusters. In a partitioning cluster analysis, partitioning with the number of clusters that are actually present in an examined dataset or at least with an approximation of it results in good quality of clustering. For this reason, the quality of a clustering is checked via a process called as cluster validation. It is mainly carried out for three purposes:

a) To search the number of clusters giving the optimal clustering result for a dataset,
b) To understand which of the two or more clustering algorithms applied to the same dataset is better,
c) To decide which levels of the parameters, i.e. the amount of fuzziness, perform as the best with an examined algorithm.

Usually the validity indices are classified as 'external indices', 'internal indices' and 'relative indices' (Kovács *et al* 2005; Rendón *et a*l 2011). The external indices compare the obtained classes from a clustering session with the previously known classes (Dudoit & Fridlyand 2002). In this case, it is already known which data points belong to which clusters, and this information is used as a reference for validation of clustering quality. These indices are very useful in evaluating the success of a clustering

[1] Zeynel Cebeci
Çukurova University, Adana - Turkey
zcebeci@cukurova.edu.tr

doi: 10.17700/jai.2019.10.2.537

algorithm because the real pattern of clusters is known (Liu *et a*l 2010). On the other hand, the internal indices do not require any external information, and determine the validity of the clustering results using the analyzed data only (Thalamuthu *et al* 2005). Finally, the relative indices compare the results from the runs of one or more clustering algorithms with different input parameters on the same dataset.

The internal indices are often used to assess the clustering quality because clustering is an unsupervised learning technique, that is, it is used to determine the clustering tendencies on a dataset when its structure is unknown. In the literature, various internal validity indices have been proposed for validating partitioning clustering results. Many of them have been listed and reviewed in detail in several studies (Milligan & Cooper 1985; Halkidi *et al* 2001; Liu *et al* 2010; Rendón *et al* 2011; Charrad *et al* 2012). As given in some comparative studies (Arbelaitz *et al* 2013; Van Craenendonck & Blockeel 2015; Hämäläinen *et al* 2017), most of the existing internal indices are available to use with traditional K-means and its derivative hard clustering algorithms. Hence, they cannot be efficient in assessing the results of fuzzy clustering algorithms such as Fuzzy C-means (FCM) and its successors. In this regard, a taxonomy of the internal validity indices for hard and soft clustering can be viewed in the related literature (Bensaid *et al* 1996, Halkidi *et al* 2002).

The indices such as partition coefficient and partition entropy have been originally introduced with FCM. Later, the more efficient indices have been developed to improve the performance in finding fuzzy partitions in datasets (Wang & Zhang 2007). However, each of the cluster validation indices has a number of pros and cons because the performance can be varied depending on different factors such as shape, volume, orientation and number of the clusters in the examined datasets. Although the above mentioned factors were carried out in the most of the comparison works (Bataineh *et al* 2011; Zhou *et a*l 2014; Zhu *et al* 2019), the effect of the distances between fuzzy data points and cluster centers has not been taken into account yet. But our intuition is that the effectiveness of the fuzzy internal validity indices can be also influenced by the distances of fuzzy points to the cluster centers. So, this study aims to compare the performances of fuzzy internal validity indices using the results from FCM clustering algorithm on some simulated datasets containing different number of clusters with different distances between a fuzzy data point and the cluster centers.

2. Related Works

Let $\boldsymbol{X}$ be a numeric dataset of n data objects in the p-dimensional data space R.

$$\boldsymbol{X}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\} \subseteq \mathrm{R}^p \qquad (1)$$

In Equation 1:

n is the number of data objects in the dataset $\boldsymbol{X}$, $1 \leq n \leq \infty$
p is the number of features (or variables) which describe the data objects,
$\boldsymbol{x}_j$ is the feature vector of p-length for the data object j.

The probabilistic and possibilistic clustering algorithms partition a given dataset $\boldsymbol{X}$ into c, a predefined number of clusters through the minimization of their related objective functions with some probabilistic or possibilistic constraints. In the clustering context, clusters are mostly represented by their prototypes. A prototype is generally the center of a cluster which can be either centroids or medoids. The prototypes of clusters are provided in the prototypes matrix, $\boldsymbol{V}$.

$$\boldsymbol{V}=\{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_c\} \subseteq \mathrm{R}^\mathrm{n} \qquad (2)$$

In Equation 2:

c is the number of clusters, $1 \leq c \leq n$
$\boldsymbol{v}_i$ is the prototype vector of p-length for the cluster i.

The probabilistic and possibilistic partitioning clustering algorithms start with the initialization of a cluster prototype matrix $\boldsymbol{V}$, and updates it through the iteration steps until it is stabilized. The clustering algorithms compute the membership degrees of data objects by using some distance metrics for calculation of their proximities to the cluster centers. A distance measure, $d(\boldsymbol{x}_j,\boldsymbol{v}_i)$, represents the distance between the data object $\boldsymbol{x}_j$ and cluster prototype $\boldsymbol{v}_i$. In general, the squared Euclidean distance metric is used in most of the applications:

$$d_{euc.sq}(\boldsymbol{x}_j,\boldsymbol{v}_i) \rightarrow d^2(\boldsymbol{x}_j,\boldsymbol{v}_i) = \| \boldsymbol{x}_j-\boldsymbol{v}_i \|^2 = (\boldsymbol{x}_j-\boldsymbol{v}_i)^T (\boldsymbol{x}_j-\boldsymbol{v}_i) \qquad (3)$$

The clustering algorithms are usually run with the squared Euclidean distance norm, which induces hyper-spherical clusters. Therefore they are able find the clusters with the same shape and orientation because the norm inducing matrix is an identity matrix: $\boldsymbol{A}$=$\boldsymbol{I}$. On the other hand, the distance metrics can be also employed with a diagonal norm inducing matrix $\boldsymbol{A}=\boldsymbol{I}\ 1/\sigma_j^2$ of $n\times n$ size. This norm matrix modifies the distances depending on the direction in which the distance is measured (Timm et al, 2004; Balasko et al 2005). In this case, the squared Euclidean distance with the norm matrix $\boldsymbol{A}$ is formulated as in Equation 4.

$$d_{euc.sq}(\boldsymbol{x}_j,\boldsymbol{v}_i) \rightarrow d^2_A(\boldsymbol{x}_j,\boldsymbol{v}_i) = \| \boldsymbol{x}_j-\boldsymbol{v}_i \|^2_A = (\boldsymbol{x}_j-\boldsymbol{v}_i)^T \mathbf{A}\ (\boldsymbol{x}_j-\boldsymbol{v}_i) \qquad (4)$$

The partitioning clustering is based on the partition of the dataset $\boldsymbol{X}$ by minimizing the objective functions (J) of various clustering algorithms depending on a certain distance norm, cluster prototypes (or cluster centers) and other first-order conditions. Partitioning clustering algorithms are classified into two groups as hard and soft clustering algorithms. In hard clustering, each object in the dataset $\boldsymbol{X}$ can be a member of one and only one cluster. Contrarily, in soft clustering, an object is not only a member of a particular cluster but a member of all clusters with varying degrees of membership. In other words, an object is not forced to be a member of a specific cluster, on the contrary, it becomes a member of all of the clusters with some degrees ranging between 0 and 1. This fuzzification approach solves the membership problems arising due to the data objects close to the boundaries of neighbor clusters in the dataset $\boldsymbol{X}$.

2.1. Fuzzy C-Means Algorithm

Fuzzy C-Means (FCM) clustering algorithm was firstly studied by Dunn (1973) and generalized by Bezdek in 1974 (Bezdek 1981). Unlike K-means algorithm, a data point is not only the member of one cluster but also the member of all clusters with varying degrees of membership between 0 and 1. FCM is an iterative clustering algorithm that partitions the dataset into a predefined c clusters by minimizing the weighted within group sum of squared errors. The objective function of FCM can be expressed in Equation 5.

$$J_{FCM}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m\ d^2\left(\boldsymbol{x}_j,\boldsymbol{v}_i\right) \qquad (5)$$

In Equation 5:

$\boldsymbol{v}_i$ is the prototype (centers) vector for cluster i,
$\boldsymbol{x}_j$ is the feature vector for data object j,
$d^2(\boldsymbol{x}_j, \boldsymbol{v}_i)$ is the Euclidean distances between prototype $\boldsymbol{v}_i$ and the data object $\boldsymbol{x}_j$,
u_{ij} is the fuzzy membership degree of object j to the cluster i,
m is the parameter of fuzzy exponent.

In the objective function J_{FCM}, the fuzzifier exponent m is usually set to 2. However, it can be any positive real number: $1 \leq m \leq \infty$. The higher values of m result with fuzzier clusters while lower values of it give harder clusters. If it equals to 1, FCM becomes a hard algorithm and produces the same results with K-means clustering.

FCM must satisfy the constraints given in the formulas in 6, 7 and 8.

$$u_{ij} \in [0,1];\ \ 1 \leq i \leq c;\ 1 \leq j \leq n \qquad (6)$$

$$0 < \sum_{j=1}^{n} u_{ij} < n;\ \ 1 \leq i \leq c \qquad (7)$$

$$\sum_{i=1}^{c} u_{ij} = 1;\ \ 1 \leq j \leq n \qquad (8)$$

In FCM, membership degrees and cluster prototypes are minimized by updating them with Equation 9 and 10, respectively.

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d^2(x_j,v_i)}{d^2(x_j,v_k)}\right)^{\frac{1}{m-1}}\right]^{-1} \qquad 1 \leq i \leq c;\ 1 \leq j \leq n \qquad (9)$$

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{n} u_{ij}^m \, x_j}{\sum_{j=1}^{n} u_{ij}^m} \qquad 1 \le i \le c \tag{10}$$

In fuzzy clustering, there are two sources of the fuzziness in a clustering result. The first one is the overlapping degree of the clusters in the analyzed dataset. With Equation 9, a data point which has equal distance to the cluster centers of two overlapped clusters becomes the member of both in equal degree of membership. Secondly, when the proportions seen in Equation 9 are too high, the value becomes cramped around 1.

FCM has been a workhorse for fuzzy clustering in numerous application. However, it has been found that it is sensitive to noise and outliers in datasets. In order to avoid this problem, Krishnapuram and Keller (1993, 1996) proposed the Possibilistic C-Means (PCM) algorithm that relaxes the probabilistic constraint of FCM but it can generate coincident clusters with poor initializations. Hence, some other versions of FCM and PCM have been developed to eliminate the problem with the original PCM. Fuzzy Possibilistic C-Means (FPCM) algorithm (Pal *et al* 1997) and later the Possibilistic Fuzzy C-Means (PFCM) algorithm (Pal *et al* 2005) were proposed to overcome the noise sensitivity defect of FCM and the coincident clusters problem of PCM, and the row sum constraints problem of FPCM. The Possibilistic Clustering Algorithm (PCA) was proposed to improve FCM and PCM (Yang & Wu 2006). Wu *et al* (2010) introduced the Unsupervised Possibilistic Clustering (UPFC) algorithm as an extension of PCA. UPFC is an algorithm that tries to improve the noise sensitivity problem of FCM and the coincident clusters problem of PCM. It also has the advantage that it does not need an FCM initialization for possibilistic part of the clustering. Although several probabilistic and possibilistic algorithms are available for fuzzy clustering, the basic FCM is used in this study because the problematic factors that may affect the success of FCM are controlled with the simulation of datasets.

2.2. Internal Validity Indices for Fuzzy Clustering

Since clustering aims to maximize intra-class similarity and inter-class difference, the validity indices measure the compactness and separation of clusters after a clustering session. Compactness is a measure of how the data points in a cluster are interrelated or adherent to each other. Separation reveals how much a cluster is separated or distinct from the others. So, the low compactness and high degree of separation indicate a good quality of clustering. The internal validity indices compared in this study are described below.

Partition Coefficient (PC) can be considered as the first validity index proposed by Bezdek who also developed the basic fuzzy clustering (FCM) algorithm (Bezdek 1974a). Because it is calculated only from fuzzy membership values, PC is a computationally low-cost index as formulated in Equation 11. Although it is simple, its effectiveness is comparable to the indices PE and XB when the clusters are spherical in the dataset. It was even concluded that if the number of clusters to start an algorithm is chosen larger than the actual one, PC may be better than the index XB (Cebeci & Yildiz 2015).

$$I_{PC}(\boldsymbol{U}) = \frac{1}{n}\left(\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m\right) \tag{11}$$

An index value is computed in the range [*1/c, 1*] using the I_{PC} formula in Equation 11. The index value of $max\{I_{PC}(c_i)\}$; $2 \le i \le c_{max}$ gives the best clustering. Here, the number 2 and c_{max} respectively denote the minimum and the maximum number of clusters to start the FCM runs. c_i is the i^{th} number of clusters in this range. While a lower value of an index which is close to the lower boundary of the range indicates fuzzier clusters, hard clusters are obtained as it approaches to upper boundary, 1. If the index is equal to *1/c*, all members of the clusters have equal membership degrees ($u_{ij} = 1/c$) that indicates that there is no clustering tendency or the applied clustering algorithm fails to find the clusters in the given dataset (Halkidi *et al* 2002).

Modified Partition Coefficient (MPC) was proposed by Dave (1996) in order to reduce the monotonic decreasing tendency of PC depending on the magnitude of *c.*

$$I_{MPC}(\boldsymbol{U}) = 1 - \frac{1}{c-1}(1 - I_{PC}) \tag{12}$$

I_{MPC} index values are in the range [0, 1]. The value of $max\{I_{MPC}(c_i)\}$; $2 \le i \le c_{max}$ indicates the best clustering result.

Partition Entropy (PE) proposed by Bezdek (1974a) is a simple entropy based index which is calculated as in Equation 13.

$$I_{PE}(\boldsymbol{U}) = \frac{1}{n}\left(\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\ log_b(u_{ij})\right) \qquad (13)$$

In Equation 13, b represents a logarithm base. I_{PE} values are obtained in the range $[0,\ log_b(c)]$. Unlike *PC* index, smaller I_{PE} values show the presence of well separated clusters while cluster structures become fuzzy if I_{PE} values approach to the upper boundary of the range. An I_{PE} index value equal to $log_b(c)$ indicates that there is no clustering tendency in the dataset or the used algorithm fails to partition data completely. Therefore, $min\{I_{PE}(c_i)\}$; $2 \leq i \leq c_{max}$ is the index value giving the best quality of clustering.

The indices PC and PE show monotonic dependence on c, the number of clusters used to start FCM runs. By the number of clusters, while a peak is searched in the PC graph a pit point is searched by the number of clusters in the PE graph. Both indices are sensitive to the fuzziness parameter m. Thus, as m goes to 1, the indices give the same values for all c's. However, when m goes to ∞, both indices show a significant peak or valley at $c = 2$. Another disadvantage of both indices is that they do not take the structural information and shapes of clusters into account because the dataset $\boldsymbol{X}$ is not used in calculation of these indices.

Fukuyama & Sugeno (1989) suggested an internal index to fix the problems with the indices PE and PC. As it is seen in Equation 14, the first term of the Fukuyama-Sugeno Index (FS) is the compactness of clusters while the second term is the separation measure that indicates the distances of the cluster representatives from each other.

$$I_{FS}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\|^2 - \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\left\|\boldsymbol{v}_i - \frac{1}{c}\sum_{k=1}^{c}\boldsymbol{v}_k\right\|^2 \qquad (14)$$

In Equation 14, the first term takes into account geometry in the representation of $\boldsymbol{X}$ with respect to the prototypes in $\boldsymbol{V}$, and fuzziness provided with $\boldsymbol{U}$. The second term adds the distance of the prototypes from the overall mean and the fuzziness in each row of $\boldsymbol{U}$. Since smaller I_{FS} values indicate the presence of compact and well-separated clusters, $min\{I_{FS}(c_i)\}$; $2 \leq i \leq c_{max}$ gives the most successful clustering result. Pal and Bezdek (1995) reported that the index FS is sensitive to both low and high values of parameter m.

Xie and Beni (1991) developed the fuzzy validity index which is known as the Xie-Beni Index (XB) in Equation 15 when the parameter m is set to 2. The numerator of the equation considers the distance of objects in a cluster from their cluster centers and measures the compactness of fuzzy clustering. The denominator term represents the strength of the separation of clusters with the distance between cluster centers.

$$I_{XB}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{2}\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\|^2}{n\left(\min\limits_{i,k=1,..,c;\ i\neq k}\{\|\boldsymbol{v}_i - \boldsymbol{v}_k\|^2\}\right)} \qquad (15)$$

Smaller values of the index XB indicate more compact and well-separated clusters. However, the XB index decreases monotonically as c approaches to n. In order to eliminate this tendency, a c_{max} value is determined as the starting point of monotonic behavior, and then $min\{I_{XB}(c_i)\}$; $2 \leq i \leq c_{max}$ is used to find the best clustering result. Another disadvantage of the index XB is that it goes to infinity as m also goes to infinity.

Kwon's validity index (K) eliminates the problem of monotonic decreasing tendency that occurs due to the increase in the number of clusters for the index XB. For this purpose, Kwon (1998) added a second term to the nominator of the index XB in order to penalize high cluster numbers as seen in Equation 16.

$$I_K(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{2}\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\|^2 + \frac{1}{c}\sum_{i=1}^{c}\left\|v_i - \frac{1}{n}\sum_{l=1}^{n}\boldsymbol{x}_l\right\|^2}{\min\limits_{i\neq k}\{\|\boldsymbol{v}_i - \boldsymbol{v}_k\|^2\}} \qquad (16)$$

The optimal clustering for the index K is investigated with $min\{I_K(c_i)\}$; $2 \leq i \leq c_{max}$.

Chen & Linkens (2004) proposed the Chen-Linkens index (CL) that consists of two terms. As it is seen in Equation 17, the first term of CL formula reflects the compactness within a cluster. The second term indicates the separation degree between the clusters. The optimal clustering is obtained at the maximum value of *max*{$I_{CL}(c_i)$} ; $2 \leq i \leq c_{max}$.

$$I_{CL}(\boldsymbol{U}) = \frac{1}{n}\sum_{j=1}^{n} \max_{j}\left(u_{ij}\right) - \frac{1}{\sum_{i=1}^{c-1} i}\sum_{i=1}^{c-1}\sum_{l=i+1}^{c}\left(\frac{1}{n}\sum_{l=1}^{n} \min(u_{ij}, u_{lj})\right) \tag{17}$$

Fuzzy Hypervolume (FHV) or Gath-Geva clustering validity index is based on the hypervolume and density of clusters in a given dataset. The index FHV is formulated as in Equation 18 (Gath & Geva 1989).

$$I_{FHV}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \left(\sum_{i=1}^{c}[det(F_i)]\right)^{1/2} \tag{18}$$

In Equation 18, F_i is the fuzzy covariance matrix of the cluster *i*. It is calculated as seen in Equation 19.

$$F_i = \frac{\sum_{j=1}^{n} u_{ij}^{m}(x_j - \boldsymbol{v}_i)^T(x_j - \boldsymbol{v}_i)}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{19}$$

F_i is used as a measure of compactness. If the clusters are soft, a fuzzy clustering with lower I_{FHV} is expected. Hence, the value of *min*{$I_{FHV}(c_i)$} ; $2 \leq i \leq c_{max}$ indicates the best clustering.

Pakhira et al. (2004, 2005) proposed the validity indices that can be used in both hard and soft clustering. In order to differentiate these, Pakhira-Bandyopadhyay-Maulik (PBM) Index for fuzzy clustering is denoted as PBMF.

$$I_{PBMF}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \left(\frac{1}{c}\cdot\frac{E_1}{E_c}\cdot D_c\right)^p \tag{20}$$

In Equation 20:

$$E_1 = \sum_{j=1}^{n} u_{1j}\left\|\boldsymbol{x}_j - \boldsymbol{v}_1\right\| \tag{21}$$

$$E_c = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\left\|\boldsymbol{x}_j - \boldsymbol{v}_i\right\| \tag{22}$$

$$D_c = \max_{i,k=1,..,c;\; i\neq k}\left\|\boldsymbol{v}_i - \boldsymbol{v}_k\right\| \tag{23}$$

The authors of the index PBMF argue that the first term in Equation 20 reduces the value of index as *c* is increases. The second term is the ratio of E_1 to E_c. As formulated in Equation 21, E_1 is a constant value for a given dataset. Since the ratio decreases with an increase in *c*, the value of index value increases as E_c decreases which indicates more numbers of compact clusters. The third term, D_c with the formula in Equation 23, measures the maximum separation between two clusters over all possible pairs of clusters. It increases with the value of *c*. The power *p* controls the contrast between the different cluster configurations, and is assigned as 2 in general. The optimal clustering with I_{PBMF} is obtained at the maximum value of *max*{$I_{PBMF}(c_i)$} ; $2 \leq i \leq c_{max}$.

The soft version of the Silhouette index (FSIL) can also be used in the assessment of fuzzy clustering results with the formula given in Equation 24.

$$I_{FSIL}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \frac{\sum_{i=1}^{n}\left(u_{ij} - u_{ij'}\right)^{\alpha} s_i(c)}{\sum_{i=1}^{n}\left(u_{ij} - u_{ij'}\right)^{\alpha}} \tag{24}$$

In Equation 24, $s_i(c)$ is the silhouette value for the data point *i*. It is calculated with the formula shown in Equation 25.

$$s_i(c) = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{25}$$

In Equation 25, a_i is the average dissimilarity between the data point *i* and all of the remaining data points in the same cluster. b_i is the least mean dissimilarity between data point *i* and the data points in other clusters. The membership degrees u_{ij} and $u_{ij'}$ are the first and second largest values in the i^{th} row of the membership degrees matrix $\boldsymbol{U}$, respectively. α is a weighing coefficient, generally set to 1. The optimal clustering is proposed at *max*{$I_{SILF}(c_i)$} ; $2 \leq i \leq c_{max}$.

The Average Partition Density (APD) index, proposed by Gath and Geva (1989) is formulated in Equation 26. In the formula, $\boldsymbol{x}_j$ is the set of data points within a predefined region around the center of cluster i, which is the sum of the central members of cluster i. The best clustering is obtained with $max\{I_{APD}(c_i)\}$; $2 \leq i \leq c_{max}$.

$$I_{APD}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \frac{1}{c}\sum_{i=1}^{c}\left(\frac{\sum_{x\in X_j} u_{ij}}{\boldsymbol{v}_i}\right) \tag{26}$$

The validity-guided (re)clustering (VGC) algorithm uses cluster-validity information to guide a fuzzy (re)clustering process toward better solutions. It starts with a partition generated by a fuzzy clustering algorithm and then iteratively alters the partition by applying split-and-merge operations to form the clusters (Bensaid *et al* 1996). The authors proposed the Compactness / Separation (CS) ratio index in order to validate the results of VGC. The formula of CS index is shown in Equation 27. The value of $min\{I_{CS}(c_i)\}$; $2 \leq i \leq c_{max}$ gives the best clustering result.

$$I_{CS}(\boldsymbol{X};\boldsymbol{V},\boldsymbol{U}) = \sum_{i=1}^{c}\frac{\sum_{j=1}^{n} u_{ij}^{m} d^2(\boldsymbol{x}_j,\boldsymbol{v}_i)}{\sum_{j=1}^{n} u_{ij}\sum_{l=1}^{c}\|\boldsymbol{v}_i-\boldsymbol{v}_l\|^2} \tag{27}$$

3. Experimental Works on the Simulated Datasets

For experimental testing of the above mentioned research question, three artificial datasets having two, three and four clusters have been generated in order to validate the performances of the studied validity indices. In each of the clusters in all of the simulated datasets, five data points are created, one of which is the centroid. The non-centroid data points of the clusters are defined one unit away from the centroid. In other words, the radius of the clusters are 1. As illustrated with a red circle in Figure 1, a fuzzy data point positioned at an equal distance from the centroids of all the clusters has been added in each generated dataset.

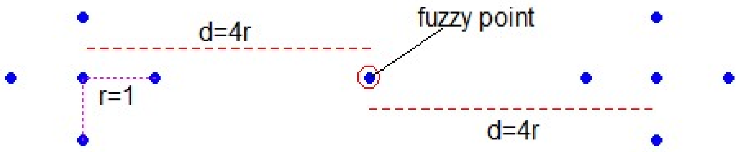

Figure 1. Pattern of the clusters and the location of fuzzy point in the simulated dataset c2r4

Five different versions of each dataset are generated by placing the fuzzy object 2 to 6 units away from the centroids of clusters. As seen in the rows of Figure 2, these datasets are named as c2r2, c3r4, etc. For instance, c3r5 stands for the dataset having three clusters in which the fuzzy point is 5 units (d = 5 × radius) away from the centroids of clusters.

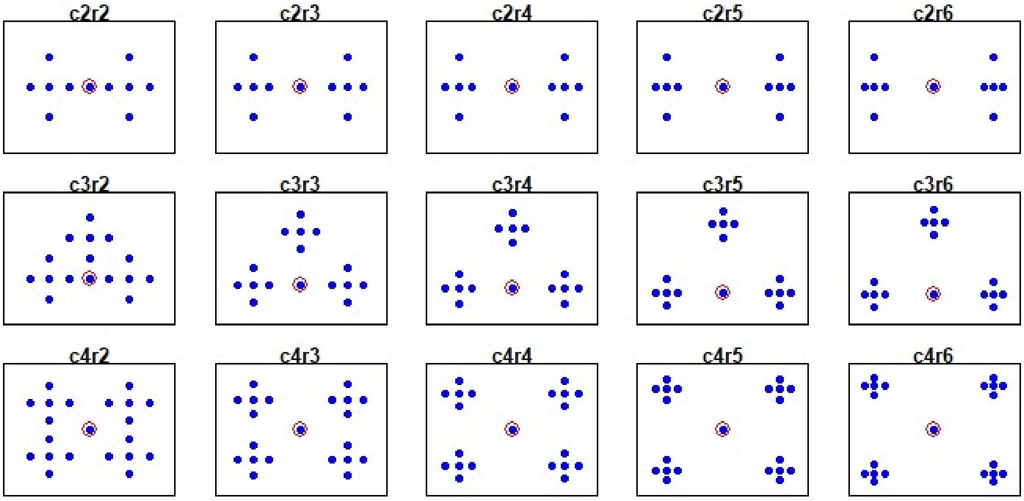

Figure 1. Scatter plots of the cluster structures in the artifical datasets

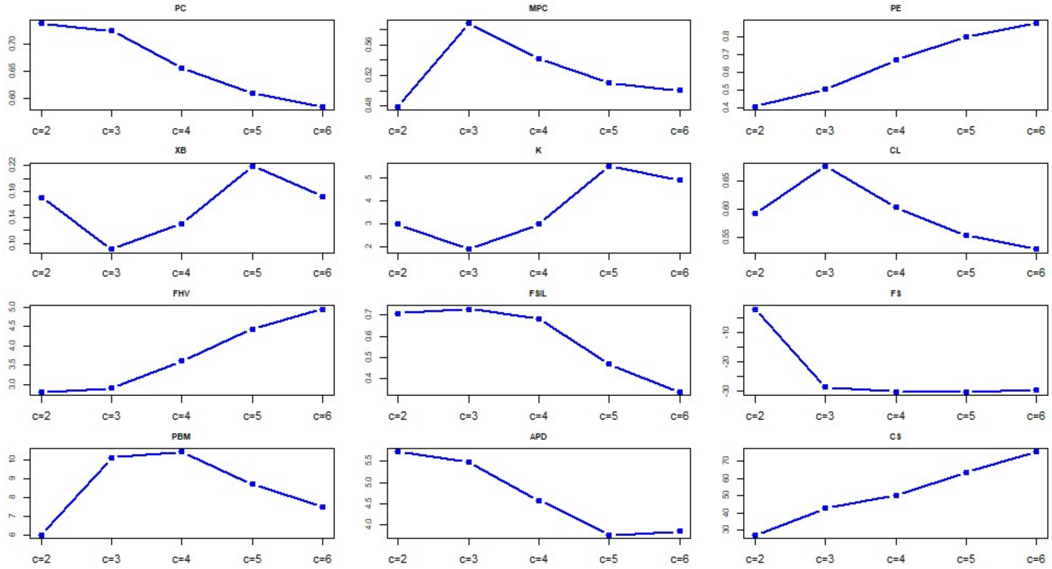

Figure 2. Changes in the values of indices by the number of clusters on the dataset c3r3

In this study, the data analysis has been performed in R environment (R Core Team, 2019). The function `fcm` of the package `ppclust` (Cebeci *et al* 2017) has been run for FCM clustering. In order to decide an optimal value of number of clusters for an examined dataset, cluster analysis should be repeated for a range of number of clusters (*c*). For this reason, the function `fcm` has been run for five levels of *c* (from 2 to 6). For each dataset, the function `fcm` has been started with the k-means++ method (Arthur & Vassilvitskii, 2007) to initialize the cluster prototype matrix $\boldsymbol{V}$. The membership matrix $\boldsymbol{U}$ has been initialized using a novel fast initialization method proposed by Cebeci (2018). The fuzzification parameter *m* has been set to 2 in all the FCM runs. The relevant functions of the package `fcvalid` (Cebeci & Cebeci 2018), downloaded from GitHub have been used for validation of fuzzy clustering result at the end of each FCM run. The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ from these successive runs of FCM, which produce the minimum objective function value have been used to evaluate the performance of validity indices. The values of indices returned by the validity functions have been checked to obtain the proposed number of clusters for each dataset, as exemplified in Figure 2 and Table 1 for the dataset containing 3 clusters in which the cluster centroids are 3 units away from the fuzzy point.

Table 1. Proposed number of clusters for the dataset c3r3

	Number of clusters examined in the FCM runs				
Index	c=2	c=3	c=4	c=5	c=6
I_{PC}	**0.740**	0.7253	0.657	0.608	0.584
I_{MPC}	0.480	**0.5880**	0.542	0.510	0.501
I_{PE}	**0.409**	0.5066	0.668	0.799	0.873
I_{XB}	0.171	**0.0908**	0.130	0.219	0.172
I_{K}	2.989	**1.8980**	2.988	5.509	4.924
I_{CL}	0.591	**0.6763**	0.603	0.553	0.528
I_{FHV}	**2.797**	2.8985	3.602	4.435	4.950
I_{FSIL}	0.708	**0.7289**	0.682	0.468	0.336
I_{FS}	-1.966	-28.8361	**-30.330**	-30.570	-29.729
I_{PBMF}	6.023	10.1041	**10.421**	8.718	7.485
I_{APD}	**5.720**	5.4909	4.582	3.761	3.848
I_{CS}	**27.252**	42.6798	50.115	63.393	75.459

As it is seen in Table 2, the indices of PE, XB, K and FSIL propose the actual number of clusters for all the datasets with two clusters. The index APD is also successful to find the actual number of clusters except the dataset c2r4. The indices PC, FHV and MPC overestimate the number of clusters for the dataset c3r6. The indices CL, FS, PBM and CS propose the number of clusters one more than the actual number of clusters in the datasets. The index FS is the worst to detect the fuzzy partitions for the dataset c2r2. The indices PE, XB, K and FSIL are stable regarding the change of distance levels of the cluster centers from the fuzzy point, and totally successful to find the actual number of clusters.

Table 2. Proposed number of clusters for the datasets with two clusters

	Distance of the fuzzy point to the cluster centers (d)				
Index	2	3	4	5	6
I_{PC}	**2**	**2**	**2**	**2**	3
I_{MPC}	**2**	**2**	**2**	3	3
I_{PE}	**2**	**2**	**2**	**2**	**2**
I_{XB}	**2**	**2**	**2**	**2**	**2**
I_{K}	**2**	**2**	**2**	**2**	**2**
I_{CL}	**2**	**2**	3	3	3
I_{FHV}	**2**	**2**	**2**	3	3
I_{FSIL}	**2**	**2**	**2**	**2**	**2**
I_{FS}	6	**2**	3	3	3
I_{PBMF}	**2**	**2**	3	3	3
I_{APD}	**2**	**2**	3	**2**	**2**
I_{CS}	**2**	**2**	3	3	3

According to the results in Table 3, the indices K and FSIL calculate the actual number of clusters at all the distance levels between the cluster centers and the fuzzy point. The indices of PE and FHV find the actual number of clusters except for the result on the dataset c3r2. All the indices except PE, K, FHV, and FSIL overestimate the number of clusters for the dataset having the clusters whose centers are 6 units away from the fuzzy point. The index APD also performs well except the datasets c3r2 and c3r6.

Table 3. Proposed number of clusters for the datasets with three clusters

	Distance of the fuzzy point to the cluster centers (d)				
Index	2	3	4	5	6
I_{PC}	2	**3**	**3**	**3**	4
I_{MPC}	**3**	**3**	**3**	4	4
I_{PE}	2	**3**	**3**	**3**	**3**
I_{XB}	**3**	**3**	**3**	**3**	4
I_{K}	**3**	**3**	**3**	**3**	**3**
I_{CL}	**3**	**3**	**3**	4	4
I_{FHV}	2	**3**	**3**	**3**	**3**
I_{FSIL}	**3**	**3**	**3**	**3**	**3**
I_{FS}	4	**3**	**3**	**4**	4
I_{PBMF}	4	**3**	**3**	4	4
I_{APD}	2	**3**	**3**	**3**	**4**
I_{CS}	2	2	**3**	4	4

For the experimental datasets having four clusters, the indices FSIL and FS have superior performance to obtain the actual number of clusters at all the levels of distances between the fuzzy point and the cluster centers in all the datasets. The indices K, CL and FHV also perform well for all the distances. The index APD underestimates the number of clusters for the distance level of 2 but overestimates for the distance levels of 5 and 6. It is also interesting that the index APD overestimates the number of clusters for all the distances in the dataset having 4 clusters when compared to the datasets having 2 and 3 clusters in which it performs well. The index CS is also totally unsuccessful to find the actual number of clusters for all the distance levels.

Table 4. Proposed number of clusters for the datasets with four clusters

	Distance of the fuzzy point to the cluster centers (d)				
Index	2	3	4	5	6
I_{PC}	2	**4**	**4**	5	5
I_{MPC}	**4**	**4**	**4**	5	5
I_{PE}	2	2	**4**	**4**	5
I_{XB}	5	5	5	5	5
I_{K}	**4**	5	**4**	**4**	**4**
I_{CL}	**4**	**4**	**4**	**4**	5
I_{FHV}	2	**4**	**4**	**4**	**4**
I_{FSIL}	**4**	**4**	**4**	**4**	**4**
I_{FS}	**4**	**4**	**4**	**4**	**4**
I_{PBMF}	**4**	**4**	**4**	5	5
I_{APD}	2	**4**	**4**	5	5
I_{CS}	2	2	5	5	5

4. Conclusions

The indices K and FSIL have successfully discovered the actual number of clusters for all levels of the distances between the fuzzy point and the cluster centers. The index FHV also performs well for the cases in which the fuzzy point is neither so close nor so far from the centers of clusters. The indices PC, MPC and PE tend to calculate the higher number of clusters than the actual ones in the cases where the fuzzy point moves away from the cluster centers. The index XB tends to overestimate the number of clusters if a dataset contains more clusters within. This result indicates that the index XB loses its validation ability for greater values of the distances. The indices FS, PBM, APD and CS return accurate results if the fuzzy point is moderately distant (d = 3-4 × avg. radius of clusters) from the cluster centers, otherwise they may not work well. These results show that these indices may not perform well if fuzzy points are too far from cluster centers.

As a general conclusion, when compared to the others, the indices FSIL, FHV and K are more stable in validating fuzzy clustering results. Additionally, the results also show that average distance between fuzzy points and cluster centers should be taken into consideration to keep the effectiveness of fuzzy validity indices more stable. A future work can enhance the information on the efficiencies of internal validity indices for fuzzy clustering results on the real datasets, sourced from various IoT- based agricultural activities. So, we plan to test the performances of the fuzzy internal validity indices on real agricultural datasets. In addition, the future work will also consider to compare the performances of the examined indices for different orientations and volumes of clusters in simulated and real datasets.

Acknowledgement

This research was supported by the grant FBA-2019-10285 from the Unit of Scientific Research Projects of Çukurova University, Adana – Turkey.

References

Arbelaitz, O, Gurrutxaga, I, Muguerza, J, Pérez, JM & Perona, I 2013 'An Extensive Comparative Study of Cluster Validity Indices'. *Pattern Recognition*, vol. 46, no. 1, pp. 243-256. doi: 10.1016/j.patcog.2012.07.021

Arthur, D & Vassilvitskii, S 2007 'K-means++: The Advantages of Careful Seeding', in Proc. of *the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027-1035.

Balasko, B, Abonyi, J & Feil, B 2005. Fuzzy clustering and data analysis toolbox. Department of Process Engineering, University of Veszprem, Veszprem. Accessed online at https://pdfs.semanticscholar.org/72f6/b22f6db1c2c0c47d8e6ead009b8c4c42bad9.pdf

Bangui, H, Ge, M, & Buhnova, B 2018 'Exploring Big Data Clustering Algorithms for Internet of Things Applications'. In Proc. of the 3rd *Int. Conf. on Internet of Things, Big Data and Security (IoTBDS 2018)*, pp. 269-276. https://doi.org/10.5220/0006773402690276

Bataineh, KM, Naji, M & Saqer, M 2011 'A Comparison Study between Various Fuzzy Clustering Algorithms'. *Jordan J. of Mechanical & Industrial Engineering*, vol. *5*, no. 4, pp. 335-343.

Bensaid, AM, Hall, L., Bezdek, JC, Clarke CP, Silbiger, ML, Arrington, JA & Murtagh, RF 1996. 'Validity-Guided (Re) Clustering with Applications to Image Segmentation'. *IEEE Transactions on Fuzzy Systems,* vol. 4, no. 2, pp. 112–123. doi: 10.1109/91.493905

Bezdek, JC 1974a 'Cluster Validity with Fuzzy Sets'. *J. Cybernet*ics, vol. 3, pp. 58–73. https://doi.org/10.1080/01969727308546047

Bezdek, JC 1974b 'Numerical Taxonomy with Fuzzy Sets'. *J. Math. Biol.,* vol. 1, pp. 57–71. https://doi.org/10.1007/BF02339490

Bezdek, JC 1981 *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York. https://doi.org/10.1007/978-1-4757-0450-1

Borgelt, C & Kruse, R 2005 'Fuzzy and Probabilistic Clustering with Shape and Size Constraints'. *Proc. of the 11th Int. Fuzzy Systems Association World Congress*, Beijing, China, pp. 945-950.

Cao, L, Zhang, X, San, X, Zhao, Y & Chen, G 2012 'Application of Fuzzy Clustering Algorithm in Precision Agriculture'. In *World Automation Congress 2012*, pp. 1-4. IEEE.

Cebeci Z & Yildiz F 2015 'Bulanık C-Ortalamalar Algoritmasının Farklı Küme Büyüklükleri için Hesaplama Performansı ve Kümeleme Geçerliliğinin Karşılaştırılması', *9. Ulusal Zootekni Bilim Kongresi,* 3-5 Eylül 2015, Konya - Türkiye. Bildiriler Kitabı, pp. 227-239. doi: https://doi.org/10.13140/RG.2.1.2909.9288

Cebeci, Z, Kavlak, AT & Yildiz, F 2017 'Validation of fuzzy and possibilistic clustering results', in Proc. of *2017 Int. Artificial Intelligence & Data Processing Symposium*, IEEE. pp. 1-7. doi: 10.1109/IDAP.2017.8090183

Cebeci, Z 2018 'Initialization of Membership Degree Matrix for Fast Convergence of Fuzzy C-Means Clustering' in Proc. of *2018 Int.Conf.on Artificial Intelligence & Data Processing,* pp. 1-5. IEEE. doi: 10.1109/IDAP.2018.8620920

Cebeci, Z & Cebeci, C 2018 'kpeaks: an R package for quick selection of k for cluster analysis' in Proc. of *2018 Int. Conf. on Artificial Intelligence & Data Processing*. pp. 1-7, IEEE. doi: 10.1109/IDAP.2018.8620896.

Charrad, M, Ghazzali, N, Boiteux, V & Niknafs, A 2014 'NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set'. *J. of Statistical Software*, vol. 61, no. 6. doi: 10.18637/jss.v061.i06

Chen, MY & Linkens, DA 2004 'Rule-Base Self-Generation and Simplification for Data-Driven Fuzzy Models'. *Fuzzy Sets and Systems*, vol. 142, no. 2, pp. 243-265. doi: 10.1016/S0165-0114(03)00160-X

Chiang, M-T. 2009 'Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads'. *J.of Classification* vol. 27, no. 3-4, pp. 3-40. doi: 10.1007/s00357-010-9049-5

Chou, C-H, Su, M-C & Lai, E 2003 'A New Cluster Validity Measure for Clusters with Different Densities', In *Proc. IASTED Int. Conf. Intell. Syst. Control*, pp. 276 -281.

Chou, C-H, Su, M-C & Lai, E 2004 'A New Cluster Validity Measure and its Application to Image Compression', *Pattern Anal. App.*vol. 7, pp. 205–220. doi: 10.1007/s10044-004-0218-1

Correa C, Valero, C, Barreiro P, Diago, MP & Tardaguila, J 2011 'A Comparison of Fuzzy Clustering Algorithms Applied to Feature Extraction on Vineyard'. In Proc. of the Conference of the Spanish Association for Artificial Intelligence, Spain, Nov 11, 2011. Accessed online at http://oa.upm.es/9246.

Dave, RN 1996, 'Validating Fuzzy Partition Obtained Through C-Shells Clustering', *Pattern Recognition Lett.* vol. 17, no. 6, pp. 613–623. doi: 10.1016/0167-8655(96)00026-8

Di Martino, F & Sessa, S 2009 'Implementation of the Extended Fuzzy C-Means Algorithm in Geographic Information Systems'. *J. of Uncertain Systems*, vol. 3, no. 4, pp. 298-306.

Dudoit, S & Fridlyand, J 2002 'A Prediction-based Resampling Method for Estimating the Number of Clusters in a Dataset'. *Genome Biology*, vol. 3, no. 7, pp. 1-21. doi: 10.1186/gb-2002-3-7-research0036

Dunn, JC 1973 'A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters'. *J. of Cybernetics*, vol.3, no.3, pp. 32-57. doi: 10.1080/01969727308546046

Fukuyama, F & Sugeno, M 1989 'A New Method of Choosing the Number of Clusters for the Fuzzy C-Means Method', In *Proc. of 5th Fuzzy Systems Symposium*, 1989, pp. 247–250.

Gath, I & Geva, AB 1989 'Unsupervised Optimal Fuzzy Clustering', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 11, no. 7, pp. 773–781. doi: 10.1109/34.192473

Halkidi, M, Batistakis, Y & Vazirgiannis, M 2002 'Clustering Validity Checking Methods: Part II', *ACM SIGMOD Record*, vol. 31, no. 3, pp. 19-27. doi: 10.1145/601858.601862.

Hämäläinen, J, Jauhiainen, S & Kärkkäinen, T 2017 'Comparison of Internal Clustering Validation Indices for Prototype-based Clustering'. *Algorithms*, vol. 10, no. 3, pp. 105. doi: 10.3390/a10030105

Kovács, F, Legány, C & Babos, A 2005 'Cluster Validity Measurement Techniques'. In Proc. of 6th *Int. Symposium of Hungarian Researchers on Computational Intelligence*, Nov 18-19, 2005, Budapest, Hungary.

Krishnapuram, R & Keller, J 1993 'A Possibilistic Approach to Clustering'. *IEEE Trans. Fuzzy Systems,* vol. 1, no. 2, pp. 98-110. doi: 10.1109/91.227387

Krishnapuram, R & Keller, J 1996 'The Possibilistic C-Means Algorithm: Insights and Recommendations'*. IEEE Trans. Fuzzy Systems*, vol. 4, no. 3, pp. 385-393. doi: 10.1109/91.531779

Kwon, SH 1998 'Cluster Validity Index for Fuzzy Clustering'. *Electron. Lett.* vol. 34, no. 22, pp. 2176–2178. doi: 10.1049/el:19981523

Liu, Y, Li, Z, Xiong, H, Gao, X & Wu, J 2010 'Understanding of Internal Clustering Validation Measures'. In *Proc. of 2010 IEEE Int. Conf. on Data Mining*, pp. 911-916. doi: 10.1109/ICDM.2010.35

Majumdar, J, Naraseeyappa, S & Ankalaki, S 2017 'Analysis of Agriculture Data Using Data Mining Techniques: Application of Big Data. *Journal of Big Data*, 4(1), 20. doi: 10.1186/s40537-017-0077-4

Marcu, I, Voicu, C, Drăgulinescu, AMC, Fratu, O, Suciu, G, Balaceanu, C & Andronache, MM 2019 'Overview of IoT Basic Platforms for Precision Agriculture'. In *Proc. of Int. Conf. on Future Access Enablers of Ubiquitous and Intelligent Infrastructures*, Springer: Cham, pp. 124-137.

Milligan, GW & Cooper, MC1985 'An Examination of Procedures for Determining the Number of Clusters in a Data Set', *Psychometrika*, vol. 50, pp. 159-179. doi: 10.1007/BF02294245

Pakhira, MK, Bandyopadhyay, S & Maulik, U 2004 'Validity Index for Crisp and Fuzzy Clusters', *Pattern Recognition* vol. 37, pp. 487–501. doi: 10.1016/j.patcog.2003.06.005

Pakhira, MK, Bandyopadhyay, S & Maulik, U 2005 'A Study of Some Fuzzy Cluster Validity Indices, Genetic Clustering and Application to Pixel Classification', *Fuzzy Sets Syst*. vol. 155, no.2, pp. 191–214. doi: 10.1016/j.fss.2005.04.009

Pal, K & Bezdek, JC 1995 'On Cluster Validity for the Fuzzy C-Means Model'. *IEEE Trans. Fuzzy Syst*., vol. 3, no. 3, pp. 370-379. doi: 10.1109/91.413225

Pal, NR, Pal, K & Bezdek JC 1997 'A Mixed C-Means Clustering Model'. In *Proc. of the Sixth IEEE Int. Conf. on Fuzzy Systems*, vol. 1, pp. 11-21. doi: 10.1109/FUZZY.1997.616338

Pal NR, Pal, K, Keller JM & Bezdek JC 2005 'A Possibilistic Fuzzy C-Means Clustering Algorithm', *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 517-530. doi: 10.1109/TFUZZ.2004.840099

R Core Team 2019 'R: A Language and Environment for Statistical Computing'. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org.

Ramesh D & Vardhan, BV 2013 'Data Mining Techniques and Applications to Agricultural Yield Data'. *Int. J. of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 9, pp. 3477-3480.

Rendón, E, Abundez, I, Arizmendi, A & Quiroz, EM 2011 'Internal Versus External Cluster Validation Indexes' *Int. J. of Computers and Communications*, vol. 5, no. 1, pp.27-34.

Thalamuthu, A, Mukhopadhyay, I, Zheng, X & Tseng, GC 2006, 'Evaluation and Comparison of Gene Clustering Methods in Microarray Analysis', *Bioinformatics*, vol. 22, no. 19, pp. 2405-2412. doi: 10.1093/bioinformatics/btl406

Tian, Z & Li, B 2015 'An Application of Fuzzy C-Means Based Clustering Technique in Smart Farming'. In *2015 Int. Conf.e on Intell. Systems Resch & Mechatronics Eng.,* Atlantis Press. doi: 10.2991/isrme-15.2015.150

Timm, H, Borgelt, C & Kruse, R 2001 'Fuzzy Cluster Analysis with Cluster Repulsion'. In *Proc. of the European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*, Tenerife, Spain. Accessed online at http://fuzzy.cs.ovgu.de/publications/doering/TimBorDoerKru01.pdf

Tiwari, M & Misra, B 2011 'Application of Cluster Analysis in Agriculture - A Review Article'. *Int. J. of Computer Applications*, vol. 36, no.4, pp. 43-47.

Van Craenendonck, T & Blockeel, H 2015 'Using Internal Validity Measures to Compare Clustering Algorithms'. In *Benelearn 2015*, Delft, NL, June 19, 2015, Poster presentations pp. 1-8. Accessed online at https://lirias.kuleuven.be/retrieve/330191

Vendrusculo, LG & Kaleita, AF 2011 'Modeling Zone Management in Precision Agriculture through Fuzzy C-Means Technique at Spatial Database'. In *2011 American Society of Agricultural and Biological Engineers.* Louisville, Kentucky, August 7-10, 2011, pp. 1. doi: 10.13031/2013.38168

Wang, W & Zhang, Y 2007 'On Fuzzy Cluster Validity Indices'. *Fuzzy Sets and Systems*, vol. 158, no. 19, pp. 2095-2117. doi: 10.1016/j.fss.2007.03.004

Wu, X, Wu, B, Sun, J & Fu, H 2010 'Unsupervised Possibilistic Fuzzy Clustering'. *J. of Information & Computational Science*, vol.7, no. 5, pp. 1075-1080.

Xie, XL & Beni, G 1991 'A Validity Measure for Fuzzy Clustering'. *IEEE Trans. Pattern Anal. Mach. Intell*. vol. 13, no. 8, pp. 841–847. doi: 10.1109/34.85677

Yang, MS & Wu, KL 2006 'Unsupervised Possibilistic Clustering'. *Pattern Recognition*, vol. 39, no. 1, pp.15-21. doi: 10.1016/j.patcog.2005.07.005

Zhang, Y, Wang, W, Zhang, X & Li, Y. 2008 'A Cluster Validity Index for Fuzzy Clustering'. *Information Sciences,* vol. 178, pp. 1205–1218. doi: 10.1016/j.patrec.2004.11.022

Zhou, K, Ding, S, Fu, C & Yang, S 2014 'Comparison and Weighted Summation Type of Fuzzy Cluster Validity Indices'. *Int. J. Computers Communications & Control*, vol. *9*, no. 3, pp. 370-378. https://doi.org/10.15837/ijccc.2014.3.237

Zhu, LF, Wang, J.S & Wang, HY 2019 'A Novel Clustering Validity Function of FCM Clustering Algorithm'. *IEEE Access*, vol. *7*, pp. 152289-152315. https://doi.org/10.1109/ACCESS.2019.2946599