Optimum Ekonomi ve Yönetim Bilimleri Dergisi, 2024, 11(1), 17-31
Optimum Journal of Economics and Management Sciences, 2024, 11(1), 17-31

Comparison of Hard and Fuzzy Clustering Techniques and Selection of Optimal Fuzzifier Parameter: An Application on Household Characteristics and Health Expenditures

Songül ÇINAROĞLU*

ABSTRACT

It is a challenging task for decision makers for finding the optimal classification pattern for the dataset obtained from national accounts, such as household budget survey (HBS) data. Fuzzy c-means (FCM) clustering, a fuzzy logic-based clustering algorithm, can be used effectively to find the proper cluster structure of given data sets under uncertainty. In this study, crisp (k-means) and fuzzy (FCM) clustering performances on grouping of households are compared while changing fuzzifier parameter for FCM. The results of the study reveal that FCM clustering performs better when compared with k-means clustering. It is found out that the optimal number of household groups is 5 and further, high cluster validity index scores are obtained when fuzzifier value is 1.5 in FCM clustering. High cluster validity index scores obtained from fuzzy Silhouette is compared to the crisp cluster validity index. The experimental results proved that fuzzy clustering superior grouping ability and it has better validity measures for grouping of households in a national dataset. It is observed that smaller fuzzifier value is a better choice to enhance fitness of fuzzy clustering. It is hoped that future experiments will compare the clustering abilities of FCM using datasets with different sizes and variables under the uncertainty conditions to determine the class boundary.

Keywords: Fuzzy c-means, Fuzzifier parameter, K-means, Silhouette

JEL Classification: I15, C38, D83

Sert ve Bulanık Kümeleme Tekniklerinin Karşılaştırılması ve Optimal Bulanıklaştırıcı Parametresinin Seçimi: Hanehalkı Özellikleri ve Sağlık Harcamaları Üzerine bir Uygulama

ÖZ

Hanehalkı bütçe anketi (HBS) verileri gibi ulusal hesaplardan elde edilen veri seti için en uygun sınıflandırma modelini bulmak karar vericiler için zorlu bir görevdir. Bulanık mantık tabanlı bir kümeleme algoritması olan bulanık c-means (FCM) kümeleme, belirsizlik altında verilen veri setlerinin uygun küme yapısını bulmak için etkili bir şekilde kullanılabilir. Bu çalışmada, FCM için bulanıklaştırıcı parametresi değiştirilirken hanehalklarının gruplandırılmasında kesin (k-ortalamalar) ve bulanık (FCM) kümeleme performansları karşılaştırılmıştır. Çalışmanın sonuçları, FCM kümelemesinin k-means kümeleme ile karşılaştırıldığında daha iyi performans gösterdiğini ortaya koymaktadır. FCM kümelemesinde en uygun hane grubu sayısının 5 olduğu ve ayrıca FCM kümelemesinde bulanıklaştırıcı değeri 1.5 olduğunda yüksek küme geçerlilik indeksi puanları elde edildiği görülmüştür. Fuzzy Silhouette den elde edilen yüksek küme geçerlilik indeks değerleri sert küme geçerlilik indeks değerleri ile karşılaştırılmıştır. Deneysel sonuçlar, bulanık kümelemenin üstün gruplama becerisine sahip olduğunu ve ulusal bir veri setinde hane halklarının gruplanması için daha iyi geçerlilik ölçütlerine sahip olduğunu kanıtlamıştır. Bulanık kümelemenin uygunluğunu artırmak için daha küçük bulanıklaştırıcı değerinin daha iyi bir seçim olduğu gözlemlenmiştir. Gelecekte yapılacak çalışmalarda sınıf sınırını belirlemek için belirsizlik koşulları altında farklı boyut ve değişkenlere sahip veri kümelerini kullanarak FCM'nin kümeleme yeteneklerinin karşılaştırılması umulmaktadır.

Anahtar Kelimeler: Fuzzy c-means, Fuzzifier parametresi, K-means, Silhouette

JEL Sınıflandırma: I15, C38, D83

Geliş Tarihi / Received: 23.03.2023 Kabul Tarihi / Accepted: 24.07.2023

* Doç. Dr., Hacettepe Üniversitesi, İİBF, Sağlık Yönetimi Bölümü, songulcinaroglu@gmail.com, ORCID:0000-0001-5699-8402.

1. INTRODUCTION

Data objects are clustered based on their similarity and difference in the clustering process (Zhou et al. 2017). Algorithms for clustering can be categorized based on partitions (Gerlhof et al. 1993), hierarchy (Guha et al. 2001), density (Hinneburg and Keim 1998), grid (Liao et al. 2004), and model. Further, clustering algorithms are divided into crisp (hard) (De Carvalho et al. 2021; Ferreira et al. 2016) and fuzzy (soft) techniques (Sert et al. 2015; Bonis and Oudot 2018). Each data object must belong to one cluster within the clustering results. Clustering algorithms have been used in various fields, including gene expression (Jothi et al. 2017), image processing (Sarkar et al. 2016), and anomaly detection (Izakian et al. 2015).

Popular partition-based clustering algorithms include k-means and fuzzy c-means (FCM). These two algorithms have both advantages and disadvantages. K-means performs faster, but it is susceptible to noises. On the contrary, FCM is a complex one and has a great ability to deal with noises. K-means have the disadvantage that the user must know in advance how many clusters to search. However, the k-means algorithm is faster than the FCM algorithm. Moreover, by using k-means, all data points are distributed equally, but not even when using FCM. This means k-means algorithm distributed the data points evenly. But, the distribution of the FCM algorithm varies (Vermurugan 2014).

FCM clustering is a fuzzy logic-based algorithm and has one of the soft computing approaches in which each vector belongs to every cluster with certain degree of membership (Goyal et al. 2019). A uniform cluster size is more likely to create partitions. The fuzzifier is a significant element for this clustering technique (Zhou et al. 2017; Schwämmle and Jensen 2010). Choosing an optimal fuzzifier parameter (m) effects clustering performances and the value for the size of "m" should not be too small or too large, since it degrades into hard k-means when “m” is equal to 1 and when m is ∞ membership degree for cluster centers is equivalent. Unsupervised learning is a form of machine learning that relies on data driven strategies since most of the parameters are formulated based on black-box modeling without prior knowledge. In the data based experiments, "m" equal to 2 is not a good choice for FCM, especially for sets of data with a wide range of cluster sizes. So, it is prudent to use k-means clustering rather than FCM for data sets that have significant uneven distribution. Further, a smaller fuzzifier value makes FCM clustering more accurate (Zhou et al. 2017).

The important features of clustering is the identification of correct number of clusters and to find the optimal partitioning method by using cluster validity index scores (Saha and Bandyopadhyay 2012). Average Silhouette width criterion, which is also referred as crisp Silhouette, is one of the popular cluster validity indices. It was invented for the assessment of non-fuzzy data partitions, but it has proven effective in evaluating fuzzy partitions as well. The Silhouette width is a measure of how similar a decision making unit is to its own group (Mohammadrezapour et al. 2020). As a particular case, the fuzzy Silhouette index includes the crisp Silhouette. Moreover, in the context of fuzzy cluster analysis counter, the fuzzy silhouette is more advantageous than its crisp counterpart, since it explicitly creates the fuzzy partition matrix generated by the clustering algorithm But, practically, it was created to improve crisp Silhouette's effectiveness in finding areas with high data density when there are overlapping clusters in the data set. (Campello and Hruschka 2006).

When large datasets, such as HBS, including socio-demographic and expenditure patterns of individuals, are considered, a better understanding of socioeconomic behaviors and the groupings of individuals in national accounts requires to find hidden patterns and to collect information (Dunn 1974). It is critical to note that, deep understanding of socio-economic behaviors of households is crucial for effective socio-economic planning (Xu et al. 2003). Clustering techniques are useful to find optimal grouping of households. However, the best clustering method to uncover latent patterns in HBS data is not well explained.

Fuzzy clustering is one of the well-known clustering methods and the selection of better fuzzifier value is controversial and there is still not one widely accepted criterion (Zhou and Yang, 2020). There exists number of studies in the literature about comparison of FCM with k-means clustering and artificial intelligence techniques by an application on telecommunication (Velmurugan 2014), energy consumption (Wei et al. 2018), wireless sensor networks (Su and Zhao 2018), etc. In FMC the fuzzifier *m* controls the amount of fuzziness of the final C-position in the FCM algorithm (Di Martino and Sessa 2022). Existing knowledge states that the structure of dataset influences the value of *m* and FCM results (Askari 2021; Karczmarek et al., 2021). Further studies is necessary to understand the effect of *m* value in fuzzy clustering results by using different datasets. The original contribution of this study is to determine the optimal value of *m* which is proposed by using HBS data.

The objective of this study is to examine and to make a comparison about the classification performances of crisp and fuzzy clustering techniques by incorporating k-means and FCM techniques on grouping of households by using HBS data. Further, optimal fuzzifier parameter will be determined and FCM clustering results will be generated by using optimal parameter values and illustrated with radar plotting of fuzzy clusters.

The remainder of the paper is organized as follows: Section 2 introduces the related works of FCM clustering, k-means clustering, cluster validity indices and describes the selection of optimal fuzzifier parameter. Section 3 provides experimental results on real world HBS data set. Finally, Section 4 presents the conclusions.

2. RELATED STUDIES

Following section provides detailed information about FCM, k-means clustering algorithms and cluster validity indexes.

2.1. FCM clustering

The FCM depends on fuzzy logic and the description of the membership function and is a sort of modified variant of k-means and the membership of a sample is generated in number of clusters (Dunn 1974; Bezdek 1981). A critical assumption of this method is that the total membership degree of each sample's overall membership degree across all clusters be equal to 1.

$$\sum_{i=1}^{c} u_{ik} = 1 \qquad \forall k = 1, \dots, n \tag{1}$$

Equation (1) presents that in FCM clustering, where "c" represents the number of clusters and u_{ik} represents the degree of sample membership in the *i*th cluster. By assuming the "n" sample and measuring the element "m" for them, the following algorithm is determined to divide the samples into a cluster with a known center:

1. First, for each sample/cluster ratio, a random membership degree is described.

2. Second, the initial membership degree is used to calculate the coordinates of the new center of the clusters, and the center of the clusters' coordinates are obtained using equation (2).

$$v_{ij} = \frac{\sum_{k=1}^{n} u_{ik}^{q} x_{kj}}{\sum_{k=1}^{n} u_{ik}^{q}} \quad (2)$$

Where v_{ij} is the *i*th changing value from the center of the *j*th cluster, u_{ik} is the degree of membership of *k*th sample to *i*th cluster, and x_{kj} is the value of the *j*th variable in the *k*th sample. q is the fuzzy amount in the *j*th variable in the *k*th sample which is known as fuzziness coefficient. In the case of q, there is no definite description, but it takes the values between 1.3 and 3 (Bezdek et al. 1984).

3. It is required to calculate the degree of membership of each sample to the center of each cluster after new cluster centers are determined. This calculation is performed by using Euclidean distance measure according to equation (3):

$$u_{ik} = \frac{(d_{ik}^{2})^{-1/q-1}}{\sum_{k=1}^{c}(d_{ik}^{2})^{-1/(q-1)}} \quad (3)$$

4. The objective function of the variable j is taken in an environment where the fuzzy coefficient q by using equation (4):

$$J = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{q} d_{ik}^{2} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{q} \|x_k - v_i\|^2 \quad (4)$$

5. Recreate calculations, until the distance between the calculated target functions is smaller than the pre-examined critical value (↖) in two successful stages, i.e. between 10^{-5} and 10^{-3}

In our case, first, three variables of c (number of classes), q (fuzziness coefficient), and ↖ (critical value) are pre-examined.

FCM clustering is different from other clustering algorithms such as k-means clustering since the fuzzifier parameter "m" and the nonzero membership degree m_{ij} are determined in here (Zhou and Yang 2020). An appropriate number of clusters "c" is determined in Fuzz c-means clustering, and then FCM clustering performs random selection of preliminary cluster centers. After that the initial cluster centers are selected. Then, the initial membership degrees of each data object x_j to cluster v_j are determined. The membership degree m_{ij} and cluster center v_i are renewed, respectively, as follows:

$$\mu_{ij} = 1/\sum_{r=1}^{c}[d(x_j, v_i)/d(x_j, v_r)]^{2/m-1} \quad (5)$$

$$v_i = \sum_{j=1}^{n} \mu_{ij}^{m} x_j / \sum_{j=1}^{n} \mu_{ij}^{m} \quad (6)$$

For each step, the objective functions FCM is calculated followingly:

$$J_{FCM}^{(c)} = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^{m} \|x_j - v_i\|^2 \quad (7)$$

Where v_i is the center of cluster V_i, and $v_i = \left(\frac{1}{n_i}\right)\sum_{x_{j\epsilon C_i}} x_j \,.\, n_i$ is the number of data objects in cluster V_i. $d(x_j, v_i)$ represents the Euclidean distance of data object x_j to cluster V_i. "m" represents the fuzzifier parameter. Determination of fuzzifier value will significantly affect the clustering results of FCM. A brief overview about the determination of optimal fuzzifier value in FCM clustering is provided in the following section (Memon 2018).

2.2. Determination of optimal fuzzifier value in FCM

The weighting exponent "m" also called as FCM's fuzzifier, controls the level of sharing among partitioned groups, and it is a critical measure. This value is one of the performance indicators of FCM clustering (Zhou et al. 2014). When the value of "m" is next to 1, it should be made clear using the membership degree definition in equation (5) that the cluster center that is closest to the point will be given significantly more weight than the others. Also, when the value of "m" is ∞, each cluster center will be take an approximately equal membership degree since

$$\lim_{m\to\infty} \mu_{ij} = \lim_{m\to\infty} \left\{1/\sum_{r=1}^{c}[(d(x_j, v_i)/d(x_j, v_r)]^{2/(m-1)}\right\} = 1/c. \quad (8)$$

In this regard, neither too small nor too large of a value for "m" will be acceptable. As a general rule, the value of fuzzifier "m" in FCM is set to 2.0, since this is equivalent to normalizing the coefficient linearly so that their sum is 1 (Zhou et al. 2017).

Table 1 represents a list of research literatures and ideas proposed for selecting fuzzifier in FCM clustering. Fuzzifier parameters are usually determined subjectively by users based on different applications, which can have a significant effect on clustering results (Zhou et al. 2017). 2 is the most widely used fuzzifier value in fuzzy applications (Pal and Bezdek 1995; Yu et al. 2004). The minimum value of "m" means that the maximum number of clusters may be calculated optimally, which is typically better than for higher "m," and so it assures that the determination of hardly discoverable clusters (Schwämmle and Jensen 2010).

Table 1: Suggestions on fuzzifier selection in FCM

Author(s)	Findings and suggestions on fuzzifier value (m)
Memon 2018; Shen et al. 2001; Yang and Nataliani 2017	[2]
Bezdek 1981; Janalipour and Mohammadzadeh 2017	[1.1, 5.0]
Ozkan and Turksen 2007	[1.4, 2.6]
Wu 2012	[1.5, 4.0]
Idri et al. 2016	[1.5, 3.5]
Chang and Cheung 1992	[1.25, 1.75]
Zhou et al. 2019	[2.5, 3.0]

2.3. k-means clustering

K-means is a kind of hard (crisp) clustering technique in which number of clusters "c" is determined at first and the initial cluster centers are then determined randomly. Further, the cluster centers are determined by equation (9) :

$$v_i = \sum_{x_j \in V_J} (x_j / n_i) \tag{9}$$

where x_j is the *j*th data object. V_i is the *i*th cluster. n_i is the number of data objects partitioned into cluster V_i . The object function is determined by using equation (10) :

$$j_{KM}^{(c)} = SSE = \sum_{i=1}^{c} \sum_{x_j \in V_i} \left\| x_j - v_i \right\|^2 \tag{10}$$

An iteration is finalized when it reaches the maximum number of repetitions (Zhou et al. 2017). Following section gives a brief overview about crisp and fuzzy Silhouette cluster validity indices.

2.4. Crisp Silhouette cluster validity index

The average Silhouette width criterion is commonly used validation measure for the clusters. To specify this criterion, consider a data object $j \in \{1,2,\dots N\}$ belonging to cluster $p \in \{1,\dots,c\}$. In the way of crisp partitions produced by prototypes-based clustering algorithm, for instance this means that object "*j*" is closer to the prototype of cluster "*p*" than to any other prototype. Fuzzy partitioning is a more general concept, this means that the membership of the *j*th object to the *p*th fuzzy cluster, μ_{pj}, is higher than the membership of this object to any other fuzzy cluster, i.e. $\mu_{pj} > \mu_{qj}$ for every $q \in \{1,\dots,c\}, q \neq p$ (Campello and Hruschka 2006).

Let the average distance of object "*j*" to all other objects belonging to cluster "*p*" be symbolized by a_{pj}. Also, let the average distance of this object to all objects belonging to another cluster "q", $q \neq p$, be called d_{qj}. Finally, let b_{pj} be the minimum d_{qj} computed over $q = 1,\dots,c, q \neq p$, which indicates the dissimilarity of object "*j*" to its closest neighboring clusters. Then, the Silhouette of object "*j"* defined as (Campello and Hruschka 2006):

$$s_j = \frac{b_{pj} - a_{pj}}{max\{a_{pj}, b_{pj}\}}, \tag{11}$$

The denominator is utilized just as a normalization expression. Obviously, the higher s_j, the better the placement of object "*j"* to cluster "p*"*. In case of "*p"* is a singleton, i.e. if it is created uniquely by object "j", then the Silhouette of this object is defined as $s_j = 0$ (Kaufman and Rousseeuw 1990). This restrains the crisp Silhouette, described as the average of s_j over $j = 1,2,\dots,N, i.e.$

$$CS = \frac{1}{N} \sum_{j=1}^{N} s_j, \tag{12}$$

to find the trivial solution $c = N$, with each object of the data set forming a cluster on its own. In this regard, the best partition is achieved when crisp Silhouette (CS) in equation (12) is used, that shows minimizing the intra-cluster distance (a_{pj}) while maximizing the inter-cluster distance (b_{pj}).

2.5. Fuzzy Silhouette cluster validity index

Utilizing the fuzzy Silhouette validity index, the fuzzy partition matrix is used explicitly. In those cases, the fuzzy partition matrix $P = [\mu_{ij}]_{cXN}$ is used only to impose on the data set a crisp partition $\tilde{P}$ =$[\tilde{\mu}_{ij}]_{cXN}$ to which the crisp Silhouette measure can be applied. Specifically, $\tilde{P}$ is such that $\tilde{\mu}_{ij} = 1$ if $i = \arg max_l \{\mu_{lj}\}$ and $\tilde{\mu}_{ij} = 0$, otherwise. As a result, as the fuzzy partition matrix "P" contains information on degrees of overlap, it may not be possible for crisp Silhouette to distinguish between overlapped data clusters. Further, data objects concentrated around cluster prototypes indicate regions of high density, while objects lying in an overlapping area indicate the opposite. Based upon this, importance will be given to the areas with high densities. In order to accomplish this, a criterion named fuzzy Silhouette (FS) is applied. It is defined as follows (Campello and Hruschka 2006).

$$FS = \frac{\sum_{j=1}^{N}(\mu_{pj}-\mu_{qj})^{\alpha} s_j}{\sum_{j=1}^{N}(\mu_{pj}-\mu_{qj})^{\alpha}}, \quad (13)$$

Where s_j is the Silhouette of object "*j*" according to equation (13), μ_{pj} and μ_{qj} are the first and second largest elements of the jth column of the fuzzy partition matrix, respectively, and $\alpha \geq 0$ is weighting coefficient. Equation (13) contains a crucial aspect that deserves particular attention. It differs from equation (12) as it is the weighted average (instead of an arithmetic mean) of the individual Silhouettes given by equation (11). Further, the weight of each term is generated by using the discrepancy between the related object's membership degrees in its first- and second-best matching fuzzy groups, subsequently. The object located near a cluster prototype is given more weight than the object located in an overlapping area (Campello and Hruschka 2006).

3. EXPERIMENTAL RESULTS

3.1. Dataset

In our case, Turkish Statistical Institute (TurkStat) HBS dataset was used to compare k-means and FCM clustering results. Data representing the socio-demographic, expenditure, income characteristics of households for the year 2015 (TurkStat 2015). The study variables include gender, age (being 65 years of age and older or not), insurance, education, marital status, number of household members, and monthly out-of-pocket (OOP) health expenditures. Basic statistics for the categorical and continuous study variables are shown in Table 2. It can be observed that, 86.8% of households have male household heads; 81.4% of household heads are under 65 years of age; 96% of household heads have health insurance; and 87.6% of them graduated from primary, secondary, and high school. It also reveals that 85.3% of household heads are married and 93.2% have less than seven household members. In addition, it is found that the median monthly OOP health care expense for the household is 22.70 TL.

Table 2: Basic statistics for study variables

Categoric variables	**N**	**%**	**Categoric variables**		**N**	**%**
Gender			**Education**			
Male	5903	86.8	Uneducated		842	12.4
Female	898	13.2	Primary & secondary & high		5959	87.6
Age_65			**Marital**			
Under 65	5537	81.4	Married		5804	85.3
65 and over	1264	18.6	Not married		997	14.7
Insurance			**Household Size**			
Yes	6532	96.0	Lower than 7		6339	93.2
No	269	4.0	Equal or higher than 7		462	6.8
Total	**6801**	**100**	**Total**		**6801**	**100**
Continuous variable	**N**	**Min.**	**Max.**	**Mean**	**Median**	**Std. Dev.**
OOP.h.exp*	6801	0.09	3746	52.93	22.70	139.96

*: Turkish liras

3.2. Classification results

The following section presents study findings obtained from k-means and FCM classification results, comparatively. First, k-means clustering results are presented that was performed by changing number of clusters. Then, FCM clustering results gathered by differing number of clusters are represented, determination of best fuzzifier parameter is also explained and presented. Then, cluster validity index scores of FCM clustering are presented, which are determined by using fuzzifier parameter value of 1.5. After that, k-means and FCM clustering results are shown and compared by using crisp and FS cluster validity index scores. Based on that, it is fund out that an optimal clustering will have a maximum Silhouette width (Rousseeuw 1987). Finally, visual representations of distribution of study variables among clusters are represented on a radar plot.

3.3. k-means clustering results by changing number of clusters

Figure 1 exhibits k-means clustering results obtained using Silhouette cluster width index scores and by changing number of clusters (k). It is observed that high Silhouette cluster validity index result is gathered when number of clusters is 5 (Silhouette index = 0.76). Further, when number of clusters determined is 30, low Silhouette cluster validity index result is obtained. So, it is observed that increase in the number of clusters decreases the classification performance results in terms of k-means classification performance scores.

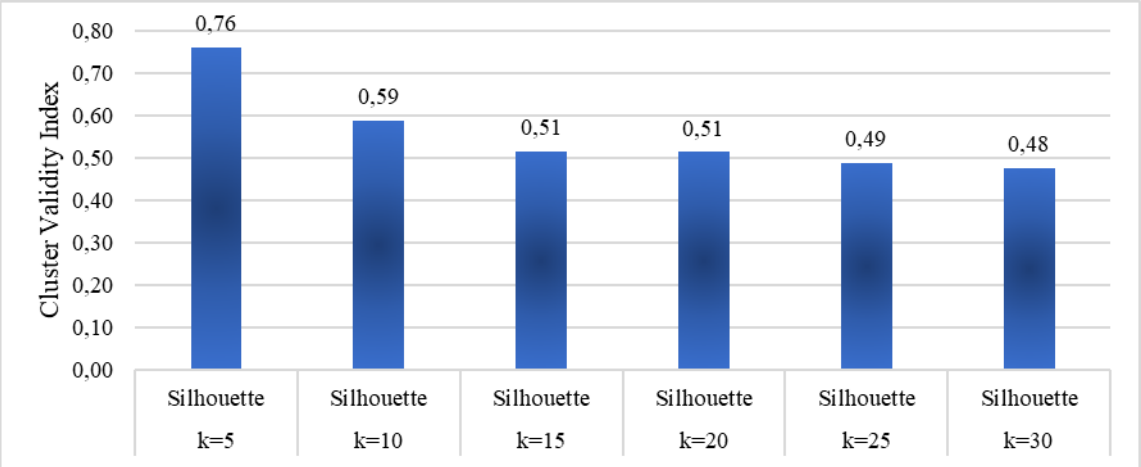

Figure 1: k-means clustering results by changing number of clusters

3.4. FCM clustering findings by differing number of clusters and finding optimal fuzzifier parameter

In this study, first FCM clustering findings examined by differing number of clusters (k) changing from 5 to 30 and changing fuzzifier parameters (m) from 1.5 to 5. High fuzzy Silhouette index (Sil.F) results obtained when number of clusters 5 and fuzzifier parameter is 1.5. Therefore, it is obvious to notice that high cluster validity index scores are gathered when lower fuzzifier parameter value is used (m=1.5). Therefore, in order to investigate how changing the number of clusters may impact the findings of the cluster validity index, the fuzzifier value m=1.5 is performed. Figure 2 displays FCM classification results generated by changing number of clusters (k) and with fuzzifier (m) parameter, simultaneously. Two cluster validity index scores are visualized in the figure, which are CS (Sil) and FS width index (Sil.F) results. It is observed that the high cluster validity index scores are achieved when number of clusters determined is 5. In addition, high cluster validity index scores are obtained when fuzzifier parameter m = 1.5 is determined. Moreover, FS cluster validity index scores are higher than crisp Silhouette index results. Further, best number of clusters for FCM clustering is found out next by using the best fuzzifier parameter and comparing crisp and fuzzy cluster validity index scores.

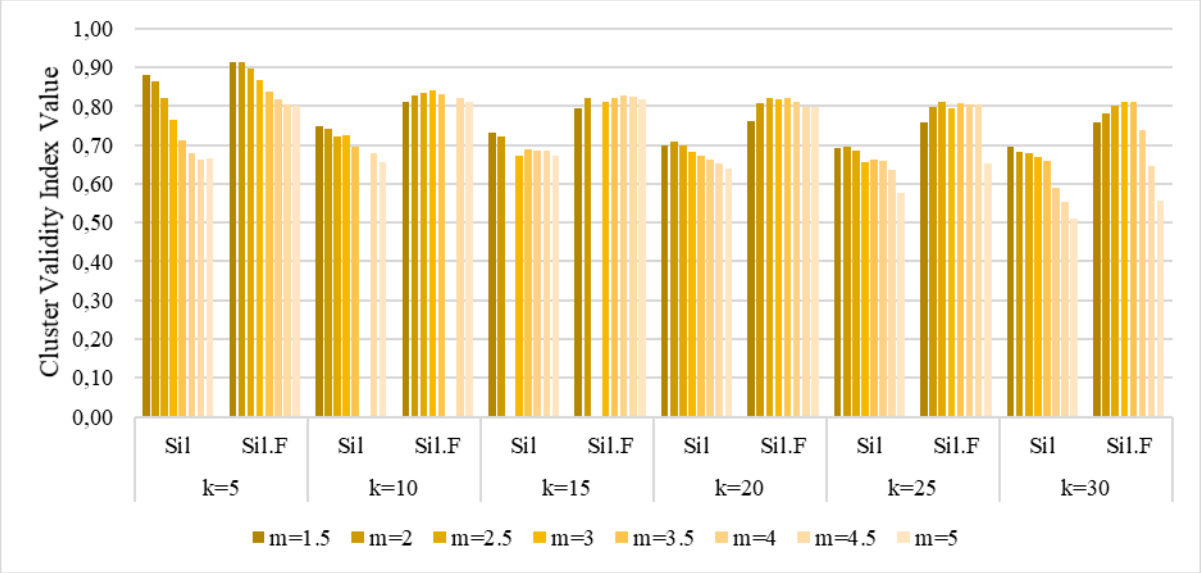

Figure 2. Changes in the fuzzifier parameter and the number of clusters in the FCM clustering

3.5. The optimal number of clusters for FCM clustering when fuzzifier parameter is 1.5

Figure 3 exhibits CS (Sil) and FS (Sil.F) width cluster validity index results obtained from FCM by changing number of clusters and determining fuzzifier parameter as m = 1.5. It is observed that the high cluster validity index scores are obtained when number of clusters is 5. In addition, higher cluster validity index scores are obtained from FS cluster validity index (Sil.F) compared with crisp Silhouette cluster validity index (Sil).

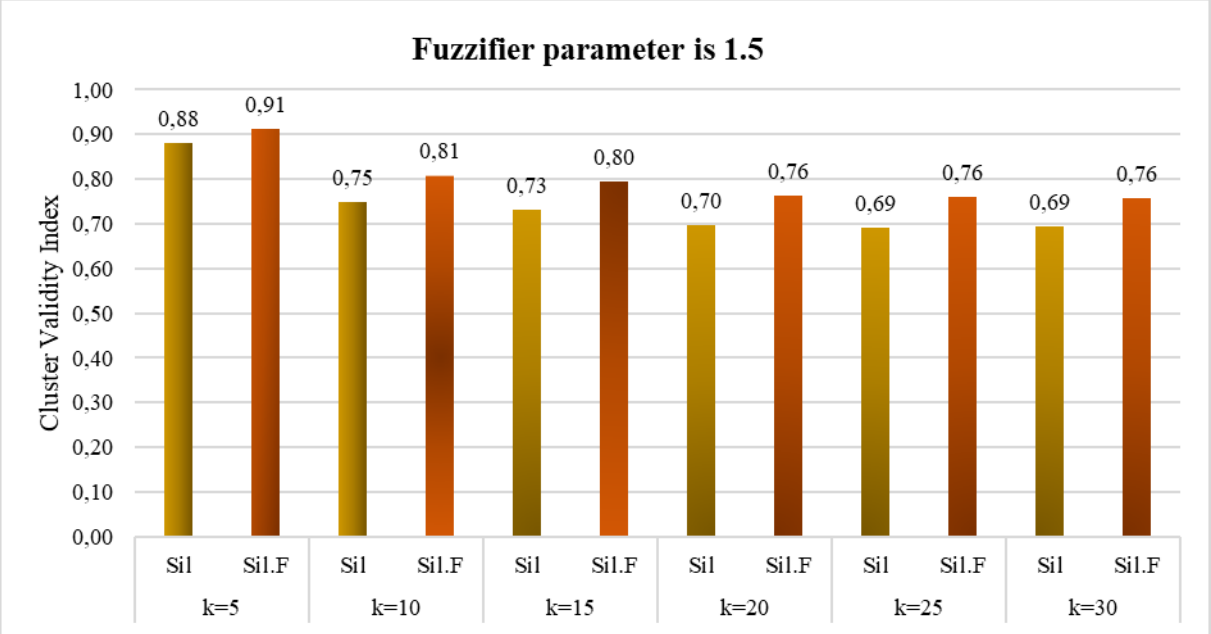

Figure 3. Best number of clusters for FCM clustering when fuzzifier parameter is 1.

3.6. Comparison of k-means and FCM clustering

Figure 4 displays the comparison of k-means and FCM clustering validity index scores obtained by using crisp Silhouette index (Sil) results. Optimal fuzzifier parameter which is 1.5 was used for FCM clustering. Comparison of crisp and fuzzy clustering results revealed that that fuzzy clustering is superior to crisp clustering technique. It is also found out that high cluster validity index score is obtained when number of clusters is 5 which is consistent with previous study findings.

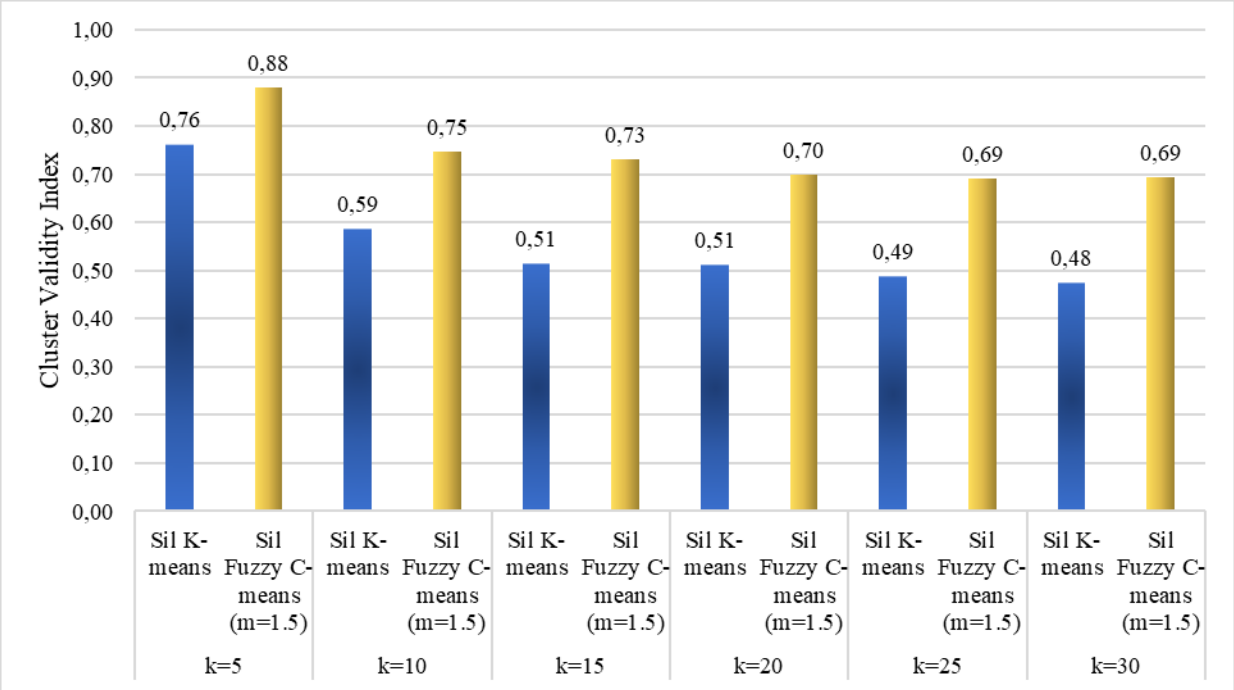

Figure 4. Comparison of k-means and FCM clustering

3.7. Radar plotting of fuzzy cluster

Figure 5 represents the visualization of radar plot created by using fuzzy clustering method. Radar plots are used to visualize profile the resulting subgroups (Zhou et al. 2019). Number of household groups in this graph is 5 and optimal fuzzifier parameter is 1.5. In fact, the five-cluster solution allows us to obtain more precise grouping results. The radar plot helps us to better understand the distribution of study variables among the five clusters. Radar plot can also be used to profiling cluster results through centroid. It is critical to note that, axis label 0 indicates mean of variable, 0.5 specifies mean plus half of standard deviation to the related variable, −0.5 indicates mean minus half of standard deviation to related variable. In our work, education discriminates households located in cluster 5, households size (hhs) differentiates households grouped in cluster 4, gender discriminates households located in cluster 1, and being equal or higher than 65 years of age or younger differentiates households located in cluster 2.

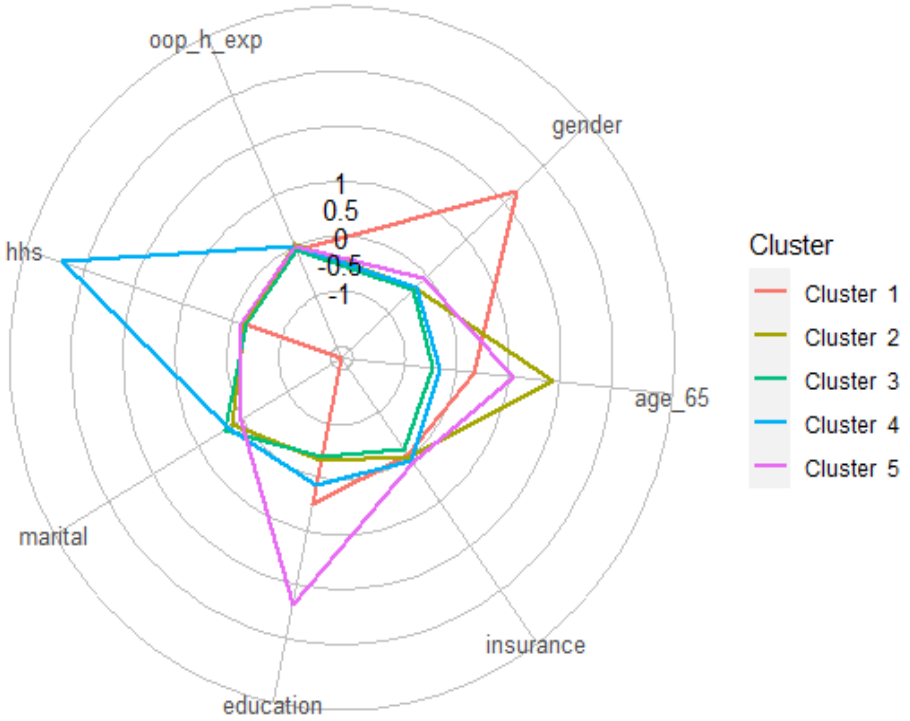

Figure 5. Radar plotting of fuzzy clusters

4. CONCLUSION

Crisp and fuzzy clustering techniques are compared in this paper to classify households based on socio-demographic and OOP health expenditure variables. The results of this study provide an obvious finding that high cluster validity index results obtained when fuzzifier parameter (m=1.5) and number of clusters (k=5) is lower. Therefore, lower fuzzifier parameter (m=1.5) is preferred to explore the effect of changing number of clusters on cluster validity index results. The fuzzy clustering model is improved in three aspects, namely fuzzifier selection, cluster validation and searching capability optimization (Zhou and Yang 2019). Literature highlights that if the value of fuzzifier m is increased then the maximum fuzzy boundary becomes wider (Salehi et al. 2021). A maximum fuzzy boundary is wider when the fuzzy parameter is greater (Pedrycz 2005). As a result of this study, it is evident that the optimal fuzzifier value is strongly dependent on the dimensions of the system and requires fine-tuning. During comparison, optimal fuzzifier parameter (m = 1.5) was determined for FCM clustering technique. Further, crisp and FS cluster validity index scores are used to compare classification performances of clustering techniques. Study findings show that FCM clustering is the best method for classifying families based on sociodemographic and OOP health expenditure factors. It is found that the best number of household group is 5 for both clustering techniques. Study findings confirms the existing knowledge by showing high cluster validity results when the fuzzifier parameter is 1.5 and the number of clusters is 5. The relationship between the fuzzifier parameter (m) and the cluster distribution is still hidden that results in the inaccuracy of the rule transformation (Huang et al., 2012). Comparing hard and fuzzy clustering techniques and choosing the best fuzzifier parameter for FCM clustering are the primary driving forces behind this study. It is highly advisable for future studies to provide a deep focus to examine the effect of changing fuzzifier parameters and number of clusters on fuzzy clustering results, simultaneously. To force the division of a data set into more clusters than it actually has,

additional applications are required to count the number of clusters with more iterations (Yang and Nataliani 2017). It is already known from the literatures that a smaller fuzzifier value is preferable (Zhou et al. 2017) and optimal fuzzifier parameter obtained for FCM clustering is 1.5 which agrees with the previous works. Moreover, FS cluster validity index scores are higher than CS clustering performance results. The results of this study prove and establish that FCM has very good ability for grouping of households in official datasets. Further, fuzzy cluster validity index results provide pioneering classification performance scores when compared to crisp counter parts. Based on this, it is highly recommended to use fuzzy logic-based classification algorithms for future studies in approaches and parameter estimation to achieve grouping of hidden patterns in official datasets. Moreover, the extensive and diverse fuzzy clustering experiments using both synthetic and real-world datasets in the future will provide deep understanding of grouping of decision-making units in national accounts.

Araştırma ve Yayın Etiği Beyanı

Bu çalışma bilimsel araştırma ve yayın etiği kurallarına uygun olarak hazırlanmıştır. Klinik ve deneysel insan ve hayvanlar üzerinde yapılmış bir çalışma olmadığından etik kurul kararı gerekmemektedir.

Yazarların Makaleye Katkı Oranları

Çalışma Songül Çınaroğlu tarafından tasarlanmış, yazılmış ve son hali verilmiştir. Yazarın makaleye katkısı %100'dür.

Çıkar Beyanı

Çıkar çatışması bulunmamaktadır.

KAYNAKÇA

Askari, S. (2021). Fuzzy C-means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: review and development. *Expert Systems with Applications*, *165*(113856), 1-27.

Bezdek J.C. (1981). *Pattern recognition with fuzzy objective algorithms*. Plenum Press. New York.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: the fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), 191-203.

Bonis, T., & Oudot, S. (2018). A fuzzy clustering algorithm for the mode-seeking framework. *Pattern Recognition Letters*, *102*, 43-73.

Campello, R. J., & Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets Systems*, *157*(2), 2858-2875.

Chan, K. P., & Cheung, Y. S. (1992). Clustering of clusters. *Pattern Recognition*, *25*(2), 211-217.

De Carvalho, F. D. A., Lechevallier, Y., & De Melo, F. M. (2021). Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition*, *45*(1), 447-464.

Di Martino, F., & Sessa, S. (2022). A novel quantum inspired genetic algorithm to initialize cluster centers in fuzzy C-means. *Expert Systems with Applications*, *191*(116340), 1-10.

Dunn J.C. (1974). A fuzzy relative ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*(3), 32-57.

Ferreira, M. R., de Carvalho, F. D. A., & Simões, E. C. (2016). Kernel based hard clustering methods with kernelization of the metric and automatic weighting of the variables. *Pattern Recognition*, *51*, 310-321.

Gerlhof, C., Kemper, A., Kilger, C., & Moerkotte, G. (1993). *Partition-based clustering in object bases: from theory to practice. foundations of data organization and algorithms*, 4th International Conference, FODO'93, Chicago, Illinois, USA, October 13-15.

Goyal, M. K., Shivam, G., & Sarma, A. K. (2019). Spatial homogeneity of extreme precipitation indices using fuzzy clustering over northeast India. *Natural Hazards*, *98*(4), 559–574.

Guha, S., Rastogi, R., & Shim, K. (2001). CURE: an efficient clustering algorithm for large databases. *Information Systems*, *26*(1), 35-58.

Hinneburg, A., & Keim, D. A. (1998). *An efficient approach to clustering in large multimedia databases with noise*. In KDD. pp. 58-65.

Huang, M., Xia, Z., Wang, H., Zeng, Q., & Wang, Q. (2012). The range of the value for the fuzzifier of the fuzzy c-means algorithm. *Pattern Recognition Letters*, *33*(16), 2280-2284.

Idri, A., Hosni, M., & Abran, A. (2016). Improved estimation of software development effect using classical and fuzzy analogy ensembles. *Applied Soft Computing*, *49*, 990-1019.

Izakian, H., Pedrycz, W., & Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, *396*, 235-244.

Janalipour, M., & Mohammadzadeh, A. (2017). Evaluation of effectiveness of three fuzzy systems and three texture extraction methods for building damage detection from post-event LiDAR data. *International Journal of Digital Earth*, *11*, 1241-1268.

Jothi, R., Mohanty, S. K., & Ojha, A. (2017). DK-means: a deterministic k-means clustering algorithm for gene expression analysis. *Pattern Analysis and Applications*, *22*, 649-667.

Karczmarek, P., Kiersztyn, A., Pedrycz, W., & Czerwiński, D. (2021). Fuzzy c-means-based isolation forest. *Applied Soft Computing*, *106*(107354), 1-10.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data*, Wiley, New York.

Liao, W. K., Liu, Y., & Choudhary, A. (2004). A grid based clustering algorithm using adaptive mesh refinement. 7th Workshop on Mining Scientific and Engineering Datasets, pp.1-9.

Memon, K. H. (2018). A histogram approach for determining fuzzifier values of interval type-2 fuzzy c-means. *Expert Systems with Applications*, *91*, 27-35.

Mohammadrezapour, O., Kisi, O., & Pourahmad, F. (2020). Fuzzy c-means and k-means clustering with genetic algorithm for identification of homogenous regions of groundwater quality. *Neural Computing & Applications*, *32*, 3763-3775.

Ozkan, & I.B. Turksen, (2007). *Upper and lower values for the level of fuzziness in FCM*. In: Wang P.P., Ruan D., Kerre E.E. (eds) Fuzzy Logic. Studies in Fuzziness and Soft Computing, vol 215. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-71258-9_6.

Pal, NR & Bezdek, JC. (1995). On cluster validity for the fuzzy c-mean model. *IEEE Transactions on Fuzzy Systems*, *3*, 370-379.

Pedrycz, W. (2005). *Knowledge-based clustering: from data to information granules*. John Wiley & Sons.

Rousseeuw, PJ. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

Saha, S., & Bandyopadhyay, S. (2012) Some connectivity based cluster validity indices. *Applied Soft Computing*, *12*(5), 1555-1565.

Salehi, F., Keyvanpour, M. R., & Sharifi, A. (2021). SMKFC-ER: Semi-supervised multiple kernel fuzzy clustering based on entropy and relative entropy. *Information Sciences*, *547*, 667-688.

Sarkar, J. P., Saha, I., & Maulik, U. (2016). Rough possibilistic type-2 fuzzy c-means clustering for MR brain image segmentation. *Applied Soft Computing*, *46*, 527-536.

Schwämmle, V., & Jensen, O. N. (2010). A simple and fast method to determine the parameters for fuzzy c–means cluster analysis. *Bioinformatics*, *26*(22), 2841-2848.

Sert, S.A., Bagci, H., & Yazici, A. (2015). MOFCA: multi-objective fuzzy clustering algorithm for wireless sensor networks. *Applied Soft Computing*, *30*, 151-165.

Shen, Y., Shi, H., & Zhang, J. Q. (2001). *Improvement and optimization of a fuzzy c-means clustering algorithm, IMTC 2001*. Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Rediscovering Measurement in the Age of Informatics (Cat. No.01CH 37188), Budapest, 3, 1430-1433.

Su, S., & Zhao, S. (2018). An optimal clustering mechanism based on Fuzzy-C means for wireless sensor networks. Sustainable *Computing: Informatics and Systems*, *18*, 127-134.

Turkish Statistical Institute (TurkStat). (2015) *Household Budget Survey Data*. https://www.tuik.gov.tr/Home/Index

Velmurugan, T. (2014). Performance based analysis between k-means and fuzzy c-means clustering algorithms for connection oriented telecommunication data. *Applied Soft Computing*, *19*, 134-146.

Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., ... & Zhao, X. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews*, *82*, 1027-1047.

Wu, K. L. (2012). Analysis parameter selections for fuzzy c-means. *Pattern Recognition*, *45*(1), 407-415.

Xu, K., Evans, D. B., Kawabata, K., Zeramdini, R., Klavus, J., & Murray, C. J. (2003). Household catastrophic health expenditure: a multicountry analysis. *The Lancet*, *362*(9378), 111-117.

Yang, M. S., & Nataliani, Y. (2017). Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters. *Pattern Recognition*, *71*, pp. 45-59.

Yu, J., Cheng, Q., & Huang, H. (2004). Analysis of weighting exponent in the FCM. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), *34*, pp. 634-639.

Zhou, F., Bai, B., Wu, Y., Chen, M., Zhong, Z., Zhu, R., ... & Zhao, Y. (2019). FuzzyRadar: visualization for understanding fuzzy clusters. *Journal of Visualization*, *22*, 913-926.

Zhou, K., & Yang, S. (2019). Fuzzifier selection in fuzzy C-means from cluster size distribution perspective. *Informatica*, *30*(3), 613-628.

Zhou, K., & Yang, S. (2020). Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering. *Pattern Analysis and Applications*, *23*, 455-466.

Zhou, K., Fu, C., & Yang, S. (2014). Fuzziness parameter selection in fuzzy c-means: the perspective of cluster validation. *Science China Information Sciences*, *57*, 1-8.

Zhou, K., Yang, S., & Shao, Z. (2017). Household monthly electricity consumption pattern mining: a fuzzy clustering-based model a case study. *Journal of Cleaner Production*, *141*, 900-908.