The application of medoid-based cluster validation in desirable dietary pattern data

Weksi Budiaji¹,³*, Rifqi Ahmad Riyanto², and Suherna¹
¹Department of Agribusiness, University of Sultan Ageng Tirtayasa, Serang, 42128, Indonesia
²Department of Food Science, University of Sultan Ageng Tirtayasa, Serang, 42128, Indonesia
³Center of Excellence of Food Security, University of Sultan Ageng Tirtayasa, Serang, 42128, Indonesia
*E-mail: budiaji@untirta.ac.id

Abstract. A desirable dietary pattern (DDP) index is an index to measure the balance and variance of the nutrition intake of an individual. This index is composed of the calorie values of protein, fat, and carbohydrates. Grouping individuals based on the DDP index is required to measure and improve an individual food security state. We took 14 individual purposively as samples to fill a set of DDP questioner. They were asked about their daily food consumption. They were grouped based on the DDP variables. A 3-dimensional plot showed that there were three to four clusters. Then, medoid-based partitioning algorithms, namely partitioning around medoids (PAM) and simple k-medoids (SKM), were applied in the data set. The inputted distances were generalized distance function to vary the distance options. The cluster results were then validated by medoid-based shadow value validation. This index was comparable to the 3-dimensional plot such that four clusters were opted as the most suitable number of clusters. The barplot of the cluster results showed that cluster 1 was characterized by an abundance of fat, while cluster 2 had very sufficient carbohydrates. Cluster 3 and 4 were two clusters with opposite characteristics where the former had a shortage of protein, fat, and carbohydrates, while the latter had an abundance of them.

1. Introduction
An approach to measure an individual food variety consumption applies desirable dietary pattern (DDP) index. The maximum value of DDP index is 100 meaning that the individual food consumption is greatly vary and balance w.r.t nutrition compositions. It is calculated via a total energy based on the calorie of protein, fat, and carbohydrates intake [1]. Instead of a single index of DDP, a 3-dimensional plot of the calories is also possible to describe the grouping of individual based on the dietary patterns.

The grouping of individual is essential to measure and improve the individual food security state. It is similar to a cluster analysis or an unsupervised classification where homogeneous individuals are assigned into a group and heterogeneous ones are separated [2, 3]. In a distance-based clustering, a similarity measure of individuals can be based on their distance, the more similar individuals have the closer distance while the less similar ones have the further distance. In a partitioning algorithm, the utmost importance is the distance choice because it affects the clustering result [4]. Besides the distance choice, a validation step is also important because the partitioning result excludes pre-determined class membership, i.e. unsupervised method.
The relative, external, and internal validation methods are applicable for the clustering result to discover the best of the data structure [5, 6].

A popular internal validation index is a silhouette index [7], which becomes the general index because it offers a visual plot. The index is vis-a-vis to a centroid-shadow value [8], which offers more variety of visualization to explore the data. A medoid-shadow value, moreover, is applicable for any type of data, i.e. numerical, categorical, and mixed variable data set [9]. Thus, this article applies the medoid-shadow value in the clustering validation of the DDP data.

2. Method

The DDP data set was collected from 14 individuals purposively. They administered a set of questions of their daily consumption. The questions consist of two parts, where the first part was multiple choice answer and the other one was opened questions. The former was about the general dietary pattern awareness, while the latter was their daily dietary list.

The second part of the questions, i.e. the opened questions, was analysed by the ddp R package [10]. However, adjustments were required to comply the ddp package. Each individual food consumption has to be categorized among 217 kinds of food. Then, the dimension of the data expands $n \times 217$, where $n$ is the number of individual, and it becomes the input data for the ddp package.

The ddp package resulted in both the DDP index and calorie consumption of each individual. Although the single index of the ddp index could be obtained easily by this package, it nonetheless less depicted the individual position among others w.r.t dietary pattern. Analysing the original variables composing the ddp index, i.e. protein, fat, and carbohydrates, on the other hand, could illustrate the data structure of the individual better. These variables were also easily produced from the ddp package.

The data structure of the individual based on the protein, fat, and carbohydrates were grouped via medoid-based clustering algorithm. We applied partitioning around medoids (PAM) algorithm [11] as the most common of the medoid-based algorithms. As a comparison, we included also simple k-medoids (SKM) algorithm [12]. The detail steps for analysing the ddp data are

(i) Plot the original data set and principle component (pc).

(ii) Set the distance function. To increase the distance choices in the numerical distance options, we applied the generalized distance function (GDF) [12], which the distance between two individuals is defined as

$$d_{ij} = \left( \alpha \sum_{r=1}^{p_n} \delta_n(x_{ir}, x_{jr}) + \beta \sum_{l=1}^{p_b} \delta_b(x_{il}, x_{jl}) + \gamma \sum_{s=1}^{p_c} \delta_c(x_{is}, x_{js}) \right)^{\omega}$$

(1)

where $\alpha, \beta, \gamma$, and $\omega$ are the weights for the numerical, binary, categorical variables, and whole distance respectively, while $p_n, p_b, p_c$ are the number of numerical, binary, and categorical variables. $\delta_n, \delta_b, \delta_c$ are the numerical, binary, and categorical distances, respectively, and $\delta_n(x_{ir}, x_{jr})$ is the numerical distance between individual $i$ and individual $j$ in the variable $r$. Because the ddp data consist of numerical variables only, the binary and categorical distances are cancelled out so that equation 1 becomes

$$d_{ij} = \alpha \sum_{r=1}^{p_n} \delta_n(x_{ir}, x_{jr})$$

(2)

Thus, the distances can be Manhattan weighted by range (mrw) [13], squared Euclidean weighted by variance (sev) [14], squared Euclidean weighted by range (ser), squared Euclidean weighted by squared range (ser.2) [15], and squared Euclidean (se) [16].
(iii) Apply medoid-based algorithm, i.e. PAM and SKM, in the five distances from equation 2. We add two other common distance for numerical variables, namely (original) Euclidean and Manhattan distances.

(iv) Validate the result via the medoid-shadow value. The medoid-shadow value is obtained by

\[ msv(x) = \frac{\delta(x, m'(x)) - \delta(x, m(x))}{\delta(x, m'(x))}, \]

where the distances between individual x to the first and second closest medoid are formulated by \( \delta(x, m(x)) \) and \( \delta(x, m'(x)) \), respectively.

(v) Choose the most suitable number of cluster.
(vi) Interpret the clustering result.

The analysis of the ddp data in this article was run in R software [17] installed in an Intel i3 4GB RAM. The supporting packages are the ddp [10], kmed [18], gg3D [19], cluster [20], and geomnet [21] packages.

3. Result and Discussion

The ddp index of the 14 individuals has an average of 68 score out of 100 (Table 1). It indicates that the desirable dietary pattern of the individuals is unbalance and limited. A further exploration is necessary to analyse each dimension that contributes to the ddp index and the structure of the data. Thus, clustering of the individual is indispensable.

| Statistics | Value |
|------------|-------|
| Min        | 50.85 |
| Q1         | 59.94 |
| Median     | 67.10 |
| Mean       | 67.69 |
| Q3         | 78.36 |
| Max        | 79.25 |

3.1. Plot of the ddp data

The ddp index is composed of three dimensions, namely protein, fat, and carbohydrates. The plot of these three variables shows that there are 3 to 4 clusters (Figure 1) with 2 clusters consisting of only an individual. The principle components plot (Figure 2), which has 95% variance of the data are explained, similarly, results in 3 to 4 clusters as well.

3.2. Medoids-based clustering algorithm and validation

The first problem in the partitioning algorithms is the distance choice because a difference distance option can results in a difference result [4]. We vary the distance into seven types, namely Euclidean (eu), Manhattan (manh), Manhattan weighted by range (mrw), squared Euclidean weighted by variance (sev), squared Euclidean weighted by range (ser), squared Euclidean weighted by squared range (ser.2), and squared Euclidean (se). The data are also unstandardised for interpretation purposes.
Protein Fat Carbohydrate

Figure 1. The 3-dimensional plot of the original variables.

Figure 2. The first and second principle components plot.

Table 2. Medoid-based shadow value of the partitioning around medoids (PAM) algorithm

| k  | mrw | sev | ser | ser.2 | se  | eu  | manh |
|----|-----|-----|-----|-------|-----|-----|------|
| 2  | 0.72| 0.86| 0.87| 0.84  | 0.88| 0.74| 0.75 |
| 3  | 0.79| 0.89| 0.87| 0.88  | 0.89| 0.75| 0.71 |
| 4  | 0.93| 0.96| 0.96| 0.98* | 0.95| 0.78| 0.76 |
| 5  | 0.81| 0.92| 0.95| 0.90  | 0.93| 0.84| 0.86 |

k = the number of clusters, mrw = Manhattan weighted by range,
sev = squared Euclidean weighted by variance,
ser = squared Euclidean weighted by range,
ser.2 = squared Euclidean weighted by squared range,
se = squared Euclidean, eu = Euclidean, manh = Manhattan,
* = the highest value of the shadow value

The PAM algorithm is applied in the seven distances. Table 2 shows the medoid-based shadow value of the PAM results for the number of clusters (k) from 2 to 5 clusters, where the higher value indicates the better separation among clusters. It suggests that the best number of clusters is 4 clusters in both the sev and ser.2 distances. The SKM algorithm, on the other hand, results in 5 clusters as the best number of cluster in term of medoid-based shadow value with the se distance (Table 3).

In addition to the 3-dimensional original variables (Figure 1) and principle component plots (Figure 2), the cluster validation tables (Table 2 and 3) present the similar result. Figure 3 and 4, moreover, depicts well separated clusters indicated by a clear distinction in both the compactness and separation. While the compactness is shown by the tick lines within a cluster, the separation is pointed out by the thin lines between clusters. Thus, we prefer 4 clusters as the most suitable number of cluster because it is also practically easier to measure and monitor the individual food security state.

3.3. Cluster interpretation

Figure 5 is a barplot showing the clustering results for 4 clusters. The population means are indicated by dots and all clusters have significantly different to the population means.
Table 3. Medoid-based shadow value of the simple k-medoids (SKM) algorithm

| k | mrw | sev | ser | ser.2 | se  | eu  | manh |
|---|-----|-----|-----|-------|-----|-----|------|
| 2 | 0.72| 0.86| 0.86| 0.86  | 0.83| 0.72| 0.75 |
| 3 | 0.76| 0.89| 0.87| 0.88  | 0.85| 0.70| 0.74 |
| 4 | 0.66| 0.83| 0.88| 0.80  | 0.89| 0.78| 0.76 |
| 5 | 0.81| 0.90| 0.91| 0.90  | 0.93| 0.84| 0.86 |

k = the number of clusters, mrw = Manhattan weighted by range, sev = squared Euclidean weighted by variance, ser = squared Euclidean weighted by range, ser.2 = squared Euclidean weighted by squared range, se = squared Euclidean, eu = Euclidean, manh = Manhattan, * = the highest value of the shadow value

Figure 3. Medoid-based validation visualization (k = 4) of sev distance.

Figure 4. Medoid-based validation visualization (k = 4) of ser.2 distance.

Cluster 1 (14%): fat. It has higher fat consumption than the population means meaning that its fat calorie is assured. However, the calorie of the protein is similar to that of the population means.

Cluster 2 (7%): carbohydrates. The calorie from the carbohydrates in this cluster is the
highest. Meanwhile, the fat consumption is just average equal to the population means. This cluster consists only an individual. The individual comes from a high income family with very frequent buying ready made food.

Cluster 3 (71%): lack of calorie. It has lower calorie of protein, fat, and carbohydrates compared to the population means meaning that the members of this cluster have an unbalance and limited nutrition intake. It is surprising that this cluster consist of 71% of the respondents.

Cluster 4 (7%): abundance of calorie. This cluster is an opposite cluster of the cluster 3. While cluster 3 has lack of calorie, cluster 4 consumes much of calorie. Similar to cluster 2, it consists of an individual only. The individual comes from a high income family with less frequent drinking milk and its variance and is very frequently eating house made food.

![Cluster Analysis Diagram]

**Figure 5.** Barplot of the 4 clusters ($k = 4$).

4. Conclusion
We applied medoid-based clustering algorithms and validation in the desirable dietary pattern (DDP) data. The 3-dimensional plot of the DDP variables, pc plot, and medoid-based clustering algorithms, namely partitioning around medoids (PAM) and simple k-medoids (SKM), resulted in 4 clusters as the most suitable and practicable number of clusters. The various types of distance were provided in the clustering algorithms. Then, the medoid-based validation produced a well separated cluster, as well. The characteristics of the 4 clusters yielded, which were presented in a barplot, were cluster 1 had an abundance of fat, while cluster 2 was characterized by a very sufficient carbohydrates. Cluster 3 had a shortage of protein, fat, and carbohydrates, while the latter had an abundance of them.

**Acknowledgments**
This research was supported by LPPM Untirta under Penelitian Dasar Internal (PDI) Program.

**References**
[1] BKP Kementan 2015 *Panduan Perhitungan Pola Pangan Harapan (PPH)* (Badan Ketahanan Pangan Kementerian Pertanian RI: Jakarta)
[2] Hastie T, Tibshirani R and J. Friedman 2009 The Element of Statistical Learning: Data Mining, Inference, and Prediction (New York, USA: Springer)

[3] Duda R, Hart P and Stock D 2001 Pattern Classification (2 ed.) (New York, USA: John Wiley and Sons)

[4] Leisch F 2006 A toolbox for k-centroids cluster analysis Computational Statistics and Data Analysis 51 526-544

[5] Webb A and Copsey K 2011 Statistical Pattern Recognition (3 ed.) (West Sussex, UK: John Wiley and Sons)

[6] Theodoridis S and Koutroubas K 2008 Pattern Recognition (4 ed. (San Diego, USA: Elsevier Academic Press)

[7] Rousseeuw P 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis Journal of Computational and Applied Mathematics 20 53-65

[8] Leisch F 2010 Neighborhood graphs, stripes and shadow plots for cluster visualization Statistics and Computing 20 457-469

[9] Budiaji W 2019 Medoid-based shadow value validation and visualization International Journal of Advances in Intelligent Informatics 5 2 76-88

[10] Budiaji W 2019 ddp: desirable dietary pattern R package version 0.0.2 available at: https://cran.r-project.org/package=ddp

[11] Kaufman L and Rousseeuw P 1990 Finding Groups in Data (New York, USA: John Wiley and Sons)

[12] Budiaji W and Leisch F 2019 Simple K-Medoids Partitioning Algorithm for Mixed Variable Data Algorithms 12 177

[13] Gower J C 1971 A General Coefficient of Similarity and Some of Its Properties Biometrics 27 857–871

[14] Wishart D 2003 K-Means Clustering with Outlier Detection, Mixed Variables and Missing Values In Exploratory Data Analysis in Empirical Research, Proceedings of the 25th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Munich, 14–16 March 2001 (Springer: Berlin/Heidelberg, Germany) 216–226

[15] Podani J 2000 Introduction to Exploration of Multivariate Biological Data (Backhuys Publishers: Leiden, The Netherlands)

[16] Huang Z 1997 Clustering large datasets with mixed numeric and categorical values In Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining (Springer: Berlin, Germany) 21–34

[17] R Core Team 2019 R: A Language and Environment for Statistical Computing Vienna, Austria: R Foundation for Statistical Computing available at: https://www.r-project.org/

[18] Budiaji W 2019 kmmed: Distance-Based k-Medoids R package version 0.3.0 available at: https://cran.r-project.org/package=kmmed

[19] Acker D 2020 gg3D: 3D perspective plots for ggplot2 R package version 0.0.0.9000

[20] Maechler M, Rousseeuw P, Struyf A, Hubert M and Hornik K 2019 cluster: Cluster Analysis Basics and Extensions R package version 2.1.0

[21] Tyner S and Hofmann H 2016 geomnet: Network Visualization in the “ggplot2” Framework R package version 0.2.0 available at: https://cran.r-project.org/package=geomnet