ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

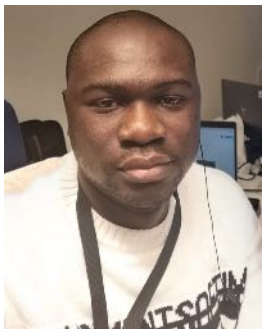
Afr. J. Food Agric. Nutr. Dev. 2021; 21(1): 17195-17212 https://doi.org/10.18697/ajfand.96.19775

K-MEANS CLUSTER ANALYSIS OF THE WEST AFRICAN SPECIES OF CEREALS BASED ON NUTRITIONAL VALUE COMPOSITION

Atsa'am DD[1]*, Oyelere SS[2], Balogun OS[3], Wario R[4] and NV Blamah[5]

Donald Douglas Atsa'am

Solomon Oyelere



Balogun Oluwafemi Samson

Ruth Wario

NV Blamah

*Corresponding author email: donatsaam@alumni.emu.edu.tr

[1]Department of Computer Science and Informatics, Faculty of Natural and Agricultural Sciences, University of the Free State, South Africa

[2]Department of Computer Science, Electrical and Space Engineering, Lulea University of Technology, Sweden

[3]School of Computing, Kuopio Campus, University of Eastern Finland, Finland

[4]Department of Computer Science and Informatics, Faculty of Natural and Agricultural Sciences, University of the Free State, South Africa

[5]Department of Computer Science, University of Jos, Jos, Nigeria

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

ABSTRACT

The *K*-means algorithm was deployed to extract clusters within the prevalent cereal foods in West Africa. The West Africa Food Composition Table (WAFCT) presents all the 76 food sources in the cereals class as a single group without considering the similarity or dissimilarity in nutritional values. Using *K*-means clustering, the Euclidean distance between nutritional values of all cereal food items were measured to generate six sub-groups based on similarity. A one-way analysis to validate the results of the extracted clusters was carried out using the mean square values. For every nutrient, the "within groups" and "between groups" values of the mean squares were examined. This was done to ascertain how similar or dissimilar data points in the same or different clusters were to each other. It was discovered that the *P* values for all "between groups" and "within groups" mean squares for every nutrient was $P < 0.01$. Additionally, it was observed that in all cases, the mean square values of the "within groups" were significantly lower than those of the "between groups". These outcomes are indications that clustering was properly done such that the variability in nutrient values for all food sources within the same clusters was significantly low, while those in different clusters were significantly high. Thus, the ultimate objective of clustering, which is to maximize intra-cluster similarity and minimize inter-cluster similarity was effectively achieved. Cluster analysis in this study showed that all food items within a particular cluster are similar to each other and dissimilar to food items in a different cluster. These findings are valuable in dietaries, food labeling, raw materials selection, public health nutrition, and food science research, when answering questions on the choice of alternative food items. Where original choices are not available or unaffordable, the clusters can be explored to select other similar options within the same cluster as the original choice.

Key words: West Africa food composition table, Cereals, Nutritional values, *K*-means clustering

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

INTRODUCTION

Data mining is the process of extracting useful information from large data sources, such as databases, data warehouses, and repositories [1,2]. The information extracted is formalized into knowledge, which can serve in decision-making. Clustering is a data mining technique undertaken to group similar data objects together into classes called clusters [3,4]. There are several methods of clustering, including partitioning, model-based, hierarchical, and fuzzy clustering.

The data mining process involves seven steps: cleaning, integration, data selection, transformation, mining, evaluation, and presentation. During the data-cleaning step, data columns that are not relevant to the task at hand are eliminated from the experimental dataset. It is possible for experimental datasets to come from multiple sources. For this reason, the integration step is required to combine such data into a single file preparatory for mining. At the selection step, relevant fields for the mining task are identified and set aside. The selected fields are then transformed into specific uniform scales or summarized forms depending on the requirement. Interesting patterns or rules are extracted from the data during the mining step, and the validity of the extracted knowledge is examined in the evaluation step. The interesting knowledge discovered in the mining process is then presented for public consumption [2,5].

The West Africa Food Composition Table (WAFCT) consists of 472 food sources defined over 28 nutrients according to the Food and Agriculture Organization of the United Nations (FAO) [6]. The food sources are drawn from nine West African countries namely, Benin, Burkina Faso, Gambia, Ghana, Guinea, Mali, Niger, Nigeria and Senegal. The 472 food sources are classified into 13 groups as: cereals and their products; starchy roots, tubers and their products; legumes and their products; vegetables and their products; fruits and their products; nuts, seeds and their products; meat, poultry and their products; eggs and their products; fish and their products; milk and their products; fats and oils; beverages; and miscellaneous. This research is limited to the cereals group since food items within this class are the most consumed in West Africa.

In the WAFCT, the cereal foods are presented as a single group consisting of 76 food sources, each of which is defined by 28 nutrients. Each food source has a nutritional value entered for each of the 28 nutrients. In the opinion of the authors, the categorization of cereal foods into a single group as previously done [6] could be altered to give rise to several sub-groups. The question this study was interested to answer is: is it possible to sub-group the 76 food sources into a number of clusters such that more closely related food items are grouped together? Cereal foods have varying amounts of chemical components and the present study researched on the possibility of sub-grouping food items with similar amounts in the same cluster. We hold the view that it is not sufficient to conclude summarily that the cereals class of food consists of 76 food sources as has been presented in the WAFCT. By measuring the distances between nutritional values, it should be possible to place food sources with shorter distances between their corresponding nutrients into same clusters. If this is successfully done, nutritionists, food industries, researchers, and patients on special diets will know other options available in situations of unaffordability or unavailability of their first choices. Therefore, the overall

objective of this study was to produce clusters that will point out the similarities that exist among food sources within the cereals group.

Clustering is an unsupervised learning activity. This means that the modeler has no idea, in advance, of what cluster membership will turn out to be. In this study, the partitioning clustering method was utilized. With this method, the dataset is subdivided into *k* groups, where *k* is the number of sub-groups specified by the modeler. In *K*-means clustering, initial clusters are arbitrarily specified after which the algorithm calculates the centers of each cluster, and then generates new clusters by assigning each object to the closest cluster center. The algorithm continues until cluster centers do not change.

Past studies conducted on cereals relating to nutritional value composition regarded all food sources under this class as a single entity. No consideration has been made to explore the possibility of establishing sub-groups within this food source based on similarity in nutrient content of group members. For instance, a previous study compared the chemical composition of cereals found in Jordan with the separate values reported by the FAO, East Asian, Moroccan, and Latin American tables [7]. In the chemical analysis, the study consistently considered cereals as one group with no mention of any sub-groups within this food class. Another study examined the differences in chemical and energy values of cereals caused by various preservation and harvest methods [8]. The research found that methods of preservation and harvest alter the chemical values of cereal grains however, the study did not mention the existence of sub-classes within the cereals food class. Another study analyzed the chemical composition of varieties of triticale, rye, and wheat cultivated in Lithuania [9]. The study found that the triticale and wheat varieties of grain exhibited similar chemical composition in comparison with the rye varieties. The research considered triticale, rye, and wheat as members of the cereals group without mentioning whether any sub-groups existed.

The West African species of cereals consists of 76 food sources, including raw grains and their products [6,10]. The WAFCT summarily presents cereals as having 76 entries without consideration as to whether there could exist sub-groups based on similarity in chemical composition. Within the cereals group, only 13 out of the 28 nutrients presented in Table 1 have no missing or all-zero values. The fact that the values of the 13 nutrients by which cereals are defined are not uniform for all 76 members, it can be possible to create different clusters that contain closely related food sources in the same sub-group. In this case, sub-group membership is determined by examining the similarity in values for each nutrient; and food sources that exhibit similarity in nutritional values can be clustered together. One of the ways of evaluating similarity or dissimilarity in numeric values is through *K*-means clustering [11-12]. As noted, the previous studies on cereals [7-9] did not contemplate that sub-groups can be created among the cereals class of food. In the same way, the WAFCT [6,10] considers cereals as one class of food without pointing out any sub-groups. The current study addresses the gap by extracting sub-groups within the cereals class of foods using the similarity and dissimilarity of chemical composition as the clustering criterion.

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

MATERIALS AND METHODS

The West African Food Composition Table (WAFCT)

Food composition tables provide information on the energy and chemical forms of nutrients that make up various food sources [13]. The information contained in food composition tables provide data that assist in formulation of dietary guidelines, estimation of amount of nutrient intake, food labeling, and public health nutrition [13-14]. According to Schönfeldt and Hall [14], food composition data are particularly important to the African continent for success of programmes aimed at combating malnutrition and food security. The WAFCT as previously reported is specifically developed to encompass food sources within the West African sub-region [6,10]. The information contained in the WAFCT was sourced from scientific papers, academic theses and dissertations, and food composition tables of Nigeria, Niger, Burkina Faso, Benin Republic, Ghana, Senegal, Guinea, Gambia, and Mali. Apart from raw food sources, the WAFCT equally consists of cooked foods categorized under appropriate food groups. The yield and nutrient retention factors were applied to calculate the nutritional values of cooked foods [15]. Categorized under 13 classes, a total of 472 food sources are contained in the WAFCT, with 28 nutrients measured in various units. The various nutrients reported in the WAFCT are presented in Table 1.

Across all food sources, values for Water, Protein, Fat, Carbohydrate, Fibre, and Ash are available. In the cases of the mineral nutrients such as calcium, iron, potassium and the rest, some food sources have no numeric values entered [6,10]. All values reported in the WAFCT are per 100 g edible portion based on average data.

Cereals

According to Alijošius *et al*. [9], cereals are cultivated worldwide in lowlands or temperate highlands environments. Cereal foods are consumed in households in several forms, while some are used as animal feed or industrial raw materials. The most popular cereals are maize, wheat, and rice while oats, barley, millets, rye, sorghum, and triticale are less popular. The research by Serna-Saldivar [16] reported that the risk of coronary heart disease and certain types of cancer are mitigated by consumption of cereal grains. Cereals constitute a source of energy, dietary fibre, and vitamin B when taken as whole grains.

The K-Means Clustering Algorithm

The *K*-means is a variant of partitioning clustering, which aims to sub-group data objects of a dataset into disjoint clusters in a way that optimizes some criteria. Consider a set of data objects $x_j \in \tilde{}^n$, $j = 1,...,N$, which are to be sub-grouped into K clusters $C = \{C_1,...C_k\}$. The optimization criterion, known as the sum-of-squared-error criterion, is given as follows [12,17].

$$f(P,M) = \sum_{i=1}^{K}\sum_{j=1}^{N} \mu_{ij} \| x_j - \mu_i \|^2$$

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

$$= \sum_{i=1}^{K} \sum_{j=1}^{N} \mu_{ij} \left(x_j - \mu_i \right)^T \left(x_j - \mu_i \right) \quad (1)$$

where

$P = \{\mu_{ij}\}$ is a partition matrix

$$\mu_{ij} = \begin{cases} 1 & \text{if } x_j \text{ belongs to cluster i} \\ 0 & \text{Otherwise} \end{cases}$$

$M = [m_1, ..., m_k]$ is the cluster means matrix

$m_i = \frac{1}{N} \sum_{j=1}^{N} \mu_{ij} x_j$ is the sample mean corresponding to the i-th cluster consisting N_i data objects. The default distance measure used in *K*-means algorithm is the Euclidean distance, and objects belonging to the same cluster should have a minimal distance as possible between them [18]. The optimal cluster is that which minimizes the criterion given in Equation (1). The *K*-means algorithm is executed by following the steps presented below:

i. Create initial *K*-partitions arbitrarily or based on advance knowledge. Evaluate the temporary cluster means matrix $M = [m_1, ..., m_k]$
ii. For each data object in the dataset, assign it to the closest cluster, C_1, that is
$x_j \in C_1, if \parallel x_j - m_l \parallel < \parallel x_j - m_i \parallel \quad j = 1, ..., N;\ i = 1, ..., K;\ i \neq l.$
iii. Re-evaluate the temporary cluster means matrix using the current partition
$m_i = \frac{1}{N_i} \sum_{x_j \in C_i} x_j$
iv. Repeat Step ii and Step iii until the means do not change [3,12,18]

Optimal Number of Clusters

In *K*-means clustering, it is required that the human expert specifies the number of clusters to be extracted in advance [4]. If the expert has prior knowledge about the dataset regarding required number of clusters, clustering can commence immediately. In situations where there is no prior knowledge, the following methods can be employed to aid the analyst in specifying the number of clusters to extract.

Elbow method: In partitioning clustering, the objective is usually to generate clusters such that the overall within-cluster sum of squares, also referred to as within-cluster variation, is minimized [19]. Using the elbow method, the *K*-means algorithm is executed on the experimental dataset using different values of *K*, such as, from 1 to 10. Then, the within-cluster sum of squares for each *K* is calculated. After this, the values of the within-cluster sum of squares for each *K* are plotted on a curve. At the location of a sharp curve in the plot, the corresponding value is considered as the optimal number of clusters.

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

Average silhouette method: This method checks to ensure that each data object is in the appropriate cluster. Using separate *K* values, the approach evaluates the mean silhouette of observations for each *K* value. When plotted; the highest point of the curve corresponds to the *K* value with the highest average silhouette, which is then selected as the optimal number of clusters [4].

Gap statistic method: This method is applicable to all the clustering techniques, not just the partitioning clustering. The total intra-cluster variations corresponding to different values of *K* for the observed data are compared with their expected values in the null reference distribution. Basically, the deviation of the observed intra-cluster variation from the expected value relating to the null hypothesis is what the gap statistic measures. The *K* value that maximizes this measure is selected as the optimal number of clusters [20-21].

Analytical Approach

Data Cleaning and Preparation

The first step in our bid to clean and prepare the data for analysis was to aggregate the data. The data was sorted and then presented in a summarized form. Secondly, fields with missing values within the dataset were eliminated, outliers identified, and noisy data smoothed out. Out of the 28 nutrients, 15 were eliminated because of missing values and 13 were retained. These include:Energy, Water, Protein, Fat, Carbohydrate, Fibre, Ash, Calcium, Iron, Vitamin D, Thiamin, Riboflavin, Niacin. Fields with non-numeric values such as 'food name' and 'food code' were also eliminated since *K*-means executes only on numeric data. Thirdly, inconsistent data were corrected and redundancy caused by data aggregation and integration resolved. Lastly, normalization, which is the process of transforming data points in all the fields to a uniform scale was performed [22]. There was a large difference between the minimum and maximum values within the data points, which required to be scaled to a uniform range. Consequently, the min-max normalization was employed and all data values were scaled to run between zero and one [1,23,24].

Cluster Tendency and Optimal Number of Clusters

Before cluster analysis is performed on any dataset, it is required that preliminary investigations are performed to determine whether the dataset is clusterable and the probable number of clusters. In this research, the optimal number of clusters to be extracted from the WAFCT dataset was determined using 3 methods, namely: Elbow, Silhouette, and gap statistic. The elbow method suggested four as the optimal number of clusters while the Silhouette method suggested three clusters. Furthermore, the Gap statistic method suggested ten clusters. It is instructive to note that each of the three methods suggested a different number of clusters to be extracted as four, three, and 10. The average of the three numbers was computed, resulting to six, which was then adopted as the optimal number of clusters.

Extraction of Clusters

The *K*-means algorithm was invoked on the Cereals data and *K* was set to six. The algorithm was executed in ten iterations to obtain the cluster membership distribution

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

(Table 2). Clusters extraction was achieved through the procedure described in the following. Using the R statistical computing software, the *K*-means algorithm was invoked on the Cereals dataset and *K* was set as six. The algorithm arbitrarily distributed the 76 food sources into six clusters and calculated the initial center of each cluster using nutrient values. After that, the algorithm calculated the Euclidean distance between each food source and each cluster center (cluster 1 to cluster 6). Next, the algorithm generated new cluster memberships by assigning each food source to the cluster with the closest center. The algorithm iteratively reshuffled the 76 food sources among the six clusters and re-computed cluster centers until the centers did not change. This was attained at the tenth iteration, giving rise to the final cluster distribution in Table 2.

RESULTS AND DISCUSSION

The detailed cluster membership for each of the 76 food sources is presented in the Appendix. The names of the food sources corresponding to each Food ID are given per cluster. It is a requirement to validate the results of any cluster analysis in order to determine how reliable the results are. In this research, three methods including Analysis of Variance (ANOVA), bar chart, and scatter plots were used to validate the extracted clusters.

Results Validation with ANOVA

Table 3 presents a one-way analysis that validates the results of the extracted clusters. For every nutrient, the "Within Groups" and "Between Groups" value of the sum of squares and mean squares are shown. The within group sum of squares or mean squares show how similar or dissimilar the data points in the same cluster are to each other. On the other hand, the between groups sum of squares or mean squares show how similar or dissimilar data points in different clusters are to each other [25]. Both the sum of squares and the mean square report the variability of data points from the mean value. According to Thinsungnoena *et al*. [26], while any of these measures could be used to report variability, it is to be noted that the sum of squares values are sensitive to the number of data points within an observation, while the mean square is not. For this reason, this study adopted the mean squares in discussing cluster validation.

It was observed in Table 3 that the *P* value for all "Between Groups" and "Within Groups" mean square for every nutrient is $P < 0.01$. In all cases, the mean square value of the "Within Groups" is significantly lower than that of the "Between Groups". These are indications that clustering has been properly done such that the variability in nutrient values for all food sources within the same clusters are significantly low, while those in different clusters are significantly high. This outcome agrees with the overall objective of clustering, which always aims to maximize intra-cluster similarity and minimize inter-cluster similarity [24,27].

Considering the "Within Groups" and "Between Groups" mean square values of Energy in Table 3, the cluster results could be interpreted to mean that, in any of the six clusters extracted, all food sources within a particular cluster exhibit a dissimilarity of an average Energy value of 0.011 to each member of that cluster. Furthermore, all food sources within separate clusters are dissimilar to each other in terms of Energy by an average

ajfand SCHOLARLY, PEER REVIEWED
AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
Volume 21 No. 1 January 2021
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

value of 1.157. It is instructive to note that 0.011 and 1.157 are relatively far apart, indicating that clustering result is valid.

Results Validation with Bar Chart

Clusters generated by the *K*-means algorithm can also be validated with the aid of a bar chart [24]. Each bar in Figure 1 represents the cluster center of a nutrient. For cluster results to be valid, it is required that no two or more bars representing the same variable in separate clusters be of the same height. That is, for example, a bar representing Protein in cluster 1 should not be of the same height with another bar representing Protein in cluster 2, cluster 3, cluster 4, cluster 5 and cluster 6. When this requirement is satisfied by the clustering result, it indicates that objects in one cluster are dissimilar to objects in another cluster. As noted, the bar chart in Figure 1 satisfies this dissimilarity requirement.

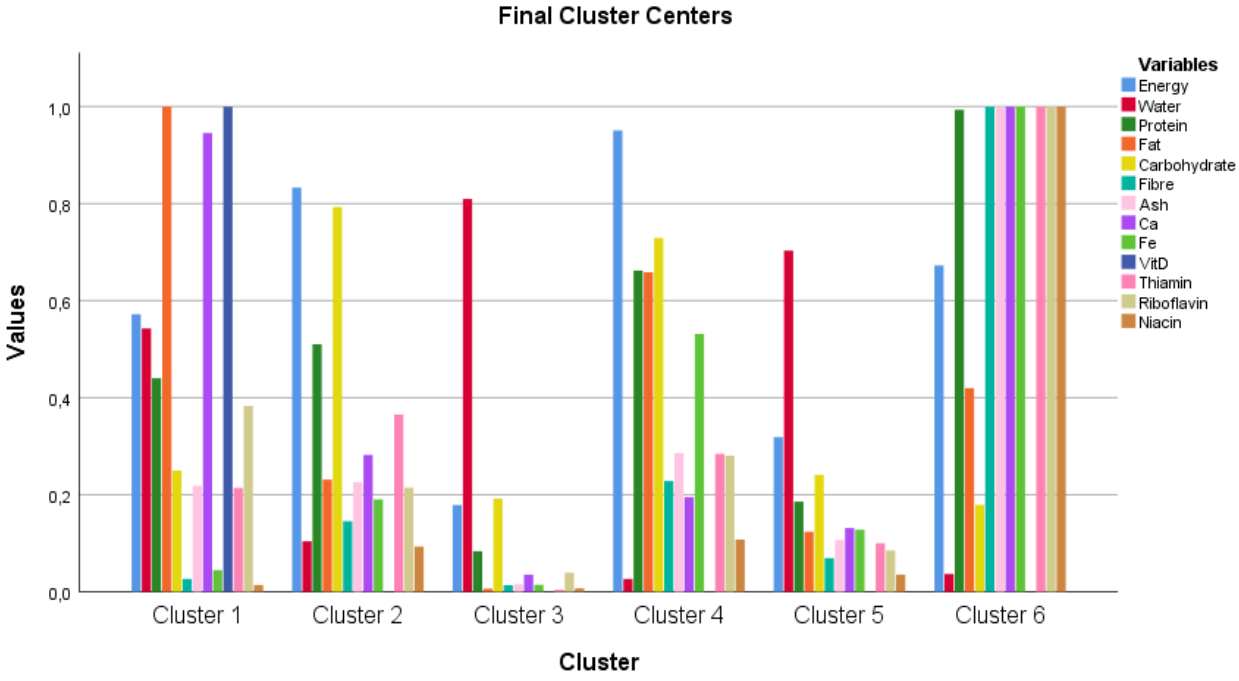


Figure 1: Bar Chart for the clusters

Results Validation with Cluster Plot

The validity of the extracted clusters was further examined with the aid of a cluster plot as shown in Figure 2. Each cluster is represented by a different colour, and it could be seen that there are no overlaps among the clusters. This is an indication that inter-clusters dissimilarity and intra-cluster similarity requirements have been satisfied in all the six clusters [24].

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
Volume 21 No. 1 January 2021
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

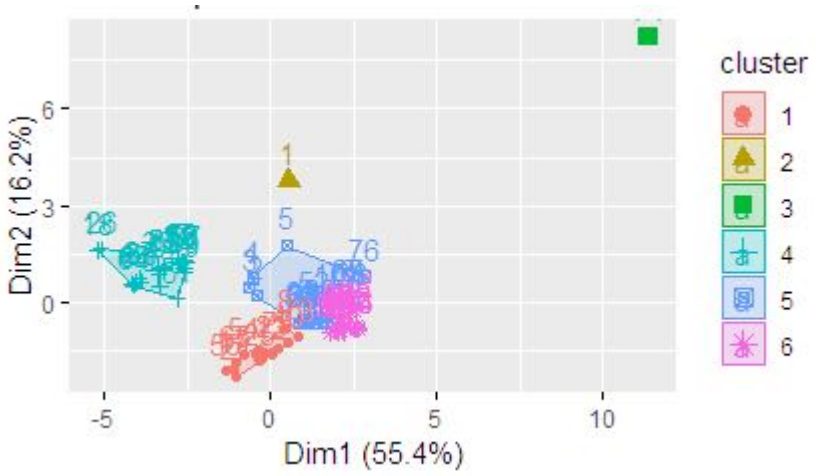


Figure 2: Cluster Plot

CONCLUSION

The West Africa Food Composition Table presents all the 76 foods in the cereals class as one big group without consideration for similarity or dissimilarity in nutritional values. This research deployed cluster analysis to extract six sub-groups within the cereals class, where food sources with similar nutritional values make up a cluster. Similarity among food items was evaluated by measuring the Euclidean distance between nutritional values that make up cereals food sources. Food items with similar nutritional values were categorized together into the same cluster. This study established that six sub-groups exist within the cereals group, proving that the West African species of cereals and their products are not just one big class of food sources. This has effectively resolved the research question in the affirmative to prove that, indeed, subgroups exist within the cereals foods. These findings are valuable in dietaries, food labeling, raw materials selection, public health nutrition, and food science and technology research. For example, a food processing industry that uses any of the food items in a particular cluster as a raw material can substitute that item for another in that same cluster. Similar options can be used as substitutes in situations of unaffordability or unavailability of original choices. In a related instance, public health nutrition can be enhanced as patients on special diets become aware of other closely related options available in order to satisfy their nutritional needs. Furthermore, in food science research experiments, food technologists can refer to the cluster results of this study to discover which food items to use as substitutes when original choices are not available. This research deployed the *K*-means clustering algorithm without consideration for other clustering algorithms. In future research, other clustering algorithms should be deployed on the cereals dataset to examine what the result would be. This study was limited to the cereals class of food, though other food classes are also found in the WAFCT. The possibility of performing cluster analysis on other food classes should be considered as part of future research activities.

Conflict of interest statement

The authors have no conflict of interest to declare.

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

Table 1: Nutrients in the WAFCT (FAO, 2012)

Nutrient	Unit
Edible portion	ratio
Energy	Kj, kcal
Water	g
Protein	g
Fat	g
Carbohydrate	g
Fibre	g
Ash	g
Calcium	mg
Iron	mg
Magnesium	mg
Phosphorous	mg
Potassium	mg
Sodium	mg
Zinc	mg
Copper	mg
Vitamin A	mcg
Retinol	mcg
Beta-carotene equivalents	mcg
Vitamin D	mcg
Vitamin E	mg
Thiamin	mg
Riboflavin	mg
Niacin	mg
Vitamin B6	mg
Folate	mcg
Vitamin B12	mcg
Vitamin C	mg

Table 2: Cluster Membership distribution

Cluster	Number of Objects
1	35
2	1
3	1
4	17
5	15
6	7
Valid	76
Missing	0

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

Table 3: One-way ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Energy	Between Groups	5.785	5	1.157	108.704	0.000
	Within Groups	0.745	70	0.011		
	Total	6.530	75			
Water	Between Groups	6.952	5	1.390	212.763	0.000
	Within Groups	0.457	70	0.007		
	Total	7.409	75			
Protein	Between Groups	3.171	5	0.634	57.160	0.000
	Within Groups	0.777	70	0.011		
	Total	3.948	75			
Fat	Between Groups	3.925	5	0.785	74.381	0.000
	Within Groups	0.739	70	0.011		
	Total	4.664	75			
Carbohydrate	Between Groups	4.987	5	0.997	89.672	0.000
	Within Groups	0.779	70	0.011		
	Total	5.766	75			
Fibre	Between Groups	1.070	5	0.214	55.522	0.000
	Within Groups	0.270	70	0.004		
	Total	1.340	75			
Ash	Between Groups	1.154	5	0.231	37.572	0.000
	Within Groups	0.430	70	0.006		
	Total	1.584	75			
Ca	Between Groups	1.630	5	0.326	22.923	0.000
	Within Groups	0.995	70	0.014		
	Total	2.625	75			
Fe	Between Groups	2.685	5	0.537	48.817	0.000
	Within Groups	0.770	70	0.011		
	Total	3.456	75			
VitD	Between Groups	0.987	5	0.197	.	.
	Within Groups	0.000	70	0.000		
	Total	0.987	75			
Thiamin	Between Groups	1.785	5	0.357	21.671	0.000
	Within Groups	1.153	70	0.016		
	Total	2.939	75			
Riboflavin	Between Groups	1.169	5	0.234	34.363	0.000
	Within Groups	0.476	70	0.007		
	Total	1.645	75			
Niacin	Between Groups	0.932	5	0.186	93.664	0.000
	Within Groups	0.139	70	0.002		
	Total	1.071	75			

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

Appendix

Food items details per Cluster

Cluster 1		
Food ID	**WAFCT Code**	**Name in English**
2	01_045	Bread/rolls, white
3	01_046	Bread, wheat, white
4	01_047	Bread, wheat, white for toasting
5	01_048	Bread, wheat, wholemeal
6	01_002	Fonio, black, whole grain, raw
8	01_050	Fonio, husked grains, raw (bran removed)
10	01_001	Fonio, white, whole grain, raw
12	01_052	Macaroni, dried
14	01_006	Maize, yellow, whole kernel, dried, raw
16	01_054	Maize, yellow, flour of whole-grain
17	01_055	Maize, yellow, grit, degermed
20	01_004	Maize, white, whole kernel, dried, raw
22	01_057	Maize, white, flour of whole grain
23	01_058	Maize, white, flour refined
24	01_059	Maize, white, flour degermed
25	01_060	Maize, white, grit, degermed
28	01_008	Maize, Gougba variety, whole kernel, dried, raw (Benin)
29	01_009	Maize, Gbaévè variety, whole kernel, dried, raw (Benin)
30	01_010	Maize, DMR-ESR-W variety, whole kernel, dried, raw (Benin)
31	01_011	Maize, POZA – RICA 7843 – SR variety, whole kernel, dried, raw (Benin)
32	01_012	Maize, TZPB-SR variety, whole kernel, dried, raw (Benin)
33	01_013	Maize, Gnonli variety, whole kernel, dried, raw (Benin)
34	01_014	Maize, combined varieties, whole kernel, dried, raw (Benin)
40	01_063	Pearl millet, flour (without bran)
50	01_034	Rice, brown, raw
52	01_065	Rice, red native, hulled, raw
54	01_067	Rice, red native, milled, raw
56	01_036	Rice, white, polished, raw
65	01_037	Rice, white, raw
67	01_039	Sorghum, whole grain, raw
69	01_041	Sorghum, whole grain, red, raw
71	01_040	Sorghum, whole grain, white, raw
73	01_072	Sorghum, flour, degermed
75	01_043	Wheat flour, white
76	01_074	Wheat, whole grains, raw
Cluster 2		
Food ID	**WAFCT Code**	**Name in English**
1	01_044	Bread, maize flour, yellow, with milk and egg

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

Cluster 3		
Food ID	**WAFCT Code**	**Name in English**
74	01_073	Wheat, bran
Cluster 4		
Food ID	**WAFCT Code**	**Name in English**
36	01_015	Millet, whole grain, raw
38	01_017	Pearl millet, whole grain, raw (with bran)
41	01_018	Pearl millet, variety ikmv 8201, whole grain, raw (Burkina Faso)
42	01_019	Pearl millet, variety ikmp 1, whole grain, raw (Burkina Faso)
43	01_020	Pearl millet, variety ikmp 2, whole grain, raw (Burkina Faso)
44	01_021	Pearl millet, variety ikmp 3, whole grain, raw (Burkina Faso)
45	01_022	Pearl millet, variety ikmp 4, whole grain, raw (Burkina Faso)
46	01_023	Pearl millet, variety ikmp 5, whole grain, raw (Burkina Faso)
47	01_031	Pearl millet, variety ikmp 13, whole gra raw (Burkina Faso)
48	01_032	Pearl millet, combined varieties, whole grain, raw (Burkina Faso)
57	01_024	Pearl millet, variety ikmp 6, whole grain raw (Burkina Faso)
58	01_025	Pearl millet, variety ikmp 7, whole grain raw (Burkina Faso)
59	01_026	Pearl millet, variety ikmp 8, whole grain raw (Burkina Faso)
60	01_027	Pearl millet, variety ikmp 9, whole grain raw (Burkina Faso)
61	01_028	Pearl millet, variety ikmp 10, whole grain, raw (Burkina Faso)
62	01_029	Pearl millet, variety ikmp 11, whole grain, raw (Burkina Faso)
63	01_030	Pearl millet, variety ikmp 12, whole grain, raw (Burkina Faso)
Cluster 5		
Food ID	**WAFCT Code**	**Name in English**
7	01_049	Fonio, black, whole grain, boiled* (without salt)
9	01_051	Fonio, husked grains, boiled* (without salt)
11	01_003	Fonio, white, whole grain, boiled* (without salt)
13	01_053	Macaroni, boiled* (without salt)
15	01_007	Maize, yellow, whole kernel, boiled* (without salt)
21	01_005	Maize, white, whole kernel, boiled* (without salt)
35	01_062	Maize, combined varieties, whole kernel, boiled* (without salt)
37	01_016	Millet, whole grain, boiled* (without salt)
39	01_033	Pearl millet, whole grain, boiled* (without salt)
49	01_064	Pearl millet, combined varieties, whole grain, boiled* (without salt) (Burkina Faso)
51	01_035	Rice, brown, boiled* (without salt)
53	01_066	Rice, red native, hulled, boiled* (without salt)
68	01_042	Sorghum, whole grain, boiled* (without salt)
70	01_070	Sorghum, whole grain, red, boiled* (without salt)
72	01_071	Sorghum, whole grain, white, boiled* (without salt)
Cluster 6		
Food ID	**WAFCT Code**	**Name in English**

ajfand SCHOLARLY, PEER REVIEWED AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT Volume 21 No. 1 January 2021 PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST ISSN 1684 5374

18	01_056	Maize, yellow, soft porridge* (without salt)
19	01_075	Maize, yellow, stiff porridge* (without salt)
26	01_061	Maize, white, soft porridge* (without salt)
27	01_076	Maize, white, stiff porridge* (without salt)
55	01_068	Rice, red native, milled, boiled* (without salt)
64	01_069	Rice, white, polished, boiled* (without salt)
66	01_038	Rice, white, boiled* (without salt)

REFERENCES

1. **Bodur EK and DD Atsa'am** Filter variable selection algorithm using risk ratios for dimensionality reduction of healthcare data for classification. *Processes,* 2019; **7**: 222. https://doi.org/10.3390/pr7040222

2. **Kantardzic M** Data Mining Concepts, Models, Methods, and Algorithms (2nd ed.). John Wiley & Sons, Hoboken. 2011.

3. **Celebi ME, Kingravi HA and PA Vela** A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications,* 2012; **40**: 200-210. https://doi.org/10.1016/j.eswa.2012.07.021

4. **Mur A, Dormido R, Duro N, Dormido-Canto S and J Vegas** Determination of the optimal number of clusters using a spectral clustering optimization. *Expert Systems With Applications* 2016; **6:** 304-314. https://doi.org/10.1016/j.eswa.2016.08.059

5. **Han J and M Kamber** Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Massachusetts. 2000.

6. **Food and Agriculture Organization of the United Nations (FAO)**. West Africa food composition table. FAO, Rome, 2012.

7. **Ereifej KI and SG Haddad** Chemical composition of selected Jordanian cereals and legumes as compared with the FAO, Moroccan, East Asian, and Latin American tables for use in the Middle East. *Trends in Food Science & Technology,* 2001; **11**: 374-378.

8. **Givens DI, Humphries DJ, Kliem KE, Kirton P and ER Deaville** Whole crop cereals 1: Effect of method of harvest and preservation on chemical composition, apparent digestibility and energy value. *Animal Feed Science and Technology,* 2009; **149**: 102-113. https://doi.org/10.1016/j.anifeedsci.2008.05.007

9. **Alijošius S Švirmickas GJ Bliznikas S Gružauskas R Šašytė V Racevičiūtė-Stupelienė A Kliševičiūtė V and A Daukšienė** Grain chemical composition of different varieties of winter cereals. *Zemdirbyste-Agriculture,* 2016; **103**: 273-280. https://doi.org10.13080/z-a.2016.103.035

10. **Stadlmayr B, Charrondiere UR and B Burlingame** Development of a regional food composition table for West Africa. *Food Chemistry,* 2013; **140**: 443-446. https://doi.org/10.1016/j.foodchem.2012.09.107

11. **Likas A, Vlassis N and JJ Verbeek** The global *k*-means clustering algorithm. *Pattern Recognition,* 2003; **3**: 451-461.

12. **Xu R and DC Wunsch** Clustering. John Wiley & Sons, Hoboken, 2009.

13. **Elmadfa I and A Meyer** Importance of food composition data to nutrition and public health. *European Journal of Chemical Nutrition,* 2010; **64**: 54-57. https://doi.org/10.1038/ejcn.2010.202

14. **Schönfeldt HC and N Hall** Capacity building in ood composition for Africa. *Food Chemistry,* 2013; **140**: 513-519. https://doi.org/10.1016/j.foodchem.2013.01.082

15. **Bergström L** Nutrient losses and gains in the preparation of foods: NLG project. *Food Chemistry,* 1996; **57**: 77-78. https://doi.org/10.1016/0308-8146(96)89017-0

16. **Serna-Saldivar SO** Cereals: Types and composition. *Reference Module in Food Science,* 2016: 718–723. https://doi.org/10.1016/B978-0-12-384947-2.00128-8

17. **Tzortzis G and A Likas** The minmax k-means clustering algorithm. *Pattern Recognition,* 2014; **47**: 2505-2516. https://doi.org/10.1016/j.patcog.2014.01.015

18. **Khan SS and A Ahmad** Cluster center initialization algorithm for K-means clustering. *Pattern Recognition Letters,* 2004; **25**: 1293-1302. https://doi.org/10.1016/j.patrec.2004.04.007

19. **Marutho D, Handaka SH and E Wijaya** The determination of cluster number at *k*-mean using elbow method and purity evaluation on headline news. *2018 International Seminar on Application for Technology of Information and Communication* (iSemantic, 2018). Pp.533-538.

20. **Tibshirani R, Walther G and T Hastie** Estimating the number of clusters in a data set via gap statistic. *Journal of the Royal Statistical Society: Series B,* 2001; **63**: 411-423.

21. **Yan M and K Ye** Determining the number of clusters using the weighted gap statistic. *Biometrics,* 2007; **63**: 1031-1037. https://doi.org/0420.2007.00784.x

22. **Kotsiantis SB, Kanellopoulos D and PE Pintelas** Data preprocessing for supervised learning. *International Journal of Computer Science,* 2006; **1**: 111-117.

23. **Atsa'am DD** Feature Selection Algorithm using Relative Odds for Data Mining Classification. In Haldorai A and & A Ramu (Eds) *Big Data Analytics for Sustainable Computing,* 2020 (pp. 81-106). Hersey, P.A: IGI Global. http://doi.org/10.4018/978-1-5225-9750-6.ch005

24. **Atsa'am DD, Wario R and EA Okpo** A new terrorism categorization based on casualties and consequences using hierarchical clustering. *Journal of Applied Security Research,* http://doi.org/10.1080/19361610.2020.1769461

25. **Mandal JK, Satapathy SC, Sanyal MK, Sarkar PP and A Mukhopadhyay** Information Systems Design and Intelligent Applications: *Proceedings of Second International Conference,* India 2015 (Vol. 2). Springer.

26. **Thinsungnoena T, Kaoungkub N, Durongdum-ronchaid P, Kerdprasopb K and N Kerdprasopb** The clustering validity with silhouette and sum of squared errors. *Learning,* 2015; **3**: 7.

27. **Forestier G, Wemmert C and P Gançarski** Semi-supervised collaborative clustering with partial background knowledge. In 2008 IEEE International Conference on Data Mining Workshops, 2008 (pp. 211-217).