http://www.ftsm.ukm.my/apjitm-2021-1002-11
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 10 No. 2, December 2021: 146 – 169
e-ISSN : 2289-2192

Received: 6 July 2021
Accepted: 21 September 2021
Published: 1 December 2021

CLUSTER ANALYSIS FOR IDENTIFYING OBESITY SUBGROUPS IN HEALTH AND NUTRITIONAL STATUS SURVEY DATA

USMAN KHALIL
OWAIS AHMED MALIK
DAPHNE TECK CHING LAI
ONG SOK KING

ABSTRACT

This study presents the discovery of meaningful patterns (groups) from the obese samples of health and nutritional survey data by applying various clustering techniques. Due to the mixed nature of the data (qualitative and quantitative variables) in the data set, the best-suited clustering techniques with appropriate dissimilarity metrics were chosen to interpret the meaningful results. The relationships between obesity and the lifestyle affecting factors like demography, socio-economic status, physical activity, and dietary behavior were assessed using four cluster techniques namely Two-Step clustering, Partition Around Medoids (PAM), Agglomerative Hierarchical clustering and, Kohonen Self Organizing Maps (SOMs). The solutions generated by these techniques were analyzed and validated by the help of cluster validity (CV) indices and later on their associations were determined with the obesity classes to discover the pattern from the obese sample. Two-Step clustering and hierarchical clustering outperformed the other applied techniques in identifying the subgroups based on the underlying hidden patterns in the data. Based on the CV indices values and the association analysis (obesity factor with the cluster solutions), two subgroups were generated and profiles of these groups have been reported. The first group belonged to the middle-aged individuals who seem to take care of their lifestyle while the other group belonged to young-aged individuals who in contrast to the first group presented a careless lifestyle factor (i.e., physical activity and dietary behavior). The salient features of these subgroups have been reported and can be proposed for the betterment in the health care industry. The research helped in identifying the interesting subsets/groups within survey data demonstrating similar characteristics and health status (i.e., prevalence of obesity with respect to lifestyle factors like physical activity, dietary behavior etc.) which will help to suggest appropriate measures/steps to be taken by the concerned departments to counter them and prevent in the population.

Keywords**:** NHANSS, Machine Learning, Two-Step, Partition Around Medoids, Agglomerative, Hierarchical, Kohonen SOMs, Clustering, Obesity.

INTRODUCTION

Non-communicable diseases (NCDs) are a leading cause of deaths in the world, resulting in approximately 71% of the deaths across the globe (Siddiqi, 2010). NCDs including obesity, cancer, heart disease, diabetes mellitus, cerebrovascular disease, hypertension, high blood pressure, high cholesterol levels, etc. are affecting the population of all ages and regions such as Asia including South-East Asia (W.H.O, 2010). Similar to the other south east asian countries, Brunei Darussalam has been facing the rise in NCDs while having the highest obesity rate in the region (MOH Brunei, 2018). Obesity is one of the major risk factors causing other non-communicable diseases such as diabetes and cardiovascular problems (Ong et al., 2017). Hence the increase in obese population has been a major concern and to overcome this problem, the management and prevention of the childhood obesity is important as these obese children may likely become obese adults in future (William Chee Fui CHONG et al., 2013). Obesity is

commonly classified using body mass index (BMI) which is a simple index of weight for height (kg/m^2) that is commonly used to classify overweightness and obesity in adults (ASEAN Secretariat, 2014; Gatta-Cherifi, 2016; MOH Brunei, 2018; W.H.O, 2010). It can be measured by anthropometric attributes that include mean weight, height, waist circumference, and body mass index. However, this classification does not completely take into consideration the population-level heterogeneity and is unable to identify the variations among obese individuals. There is evidence of the association of obesity with other factors including demographics, nutrition habits and physical activity of individuals (ASEAN Secretariat, 2014; W.H.O, 2010). Finding common patterns among obese individuals can help in devising effective interventions and treatments administered by clinicians. Thus, the focus and aim of this study was to explore and to extract the useful patterns and generate profiles of these clusters from NHANSS data through clustering techniques. This useful knowledge could be helpful for the clinicians to disseminate the proper information and prescribe the patients accordingly for the well-being of the region's population. In this research, four different clustering techniques have been applied to the survey data and clusters were identified and validated. Further, the associations between clusters and various health factors have been computed and profiling of the observed clusters was performed.

RESEARCH OBJECTIVES

The main objectives of this study were set as follows.

1. Identify the sub-groups within the obese samples from NHANSS data set by applying the clustering techniques
2. Study the patterns generated by different clustering methods for obesity as a non-communicable disease, and discover the factors affecting it
3. Interpret and profile the salient characteristics of sub-groups based on validated cluster solutions
4. Generate potential recommendations and relevant information about affecting factors of obesity to clinicians for taking preventive measures

SIGNIFICANT CONTRIBUTION

The major contribution of the study is to identify the heterogeneity within the obese group with respect to factors affecting it, such as demographic, socio-economic status, behavior characteristics like short food frequency, physical activity, and smoking habit etc. Generally, body mass index (BMI) metric is used to classify obesity in the population. It is a person's weight in kilograms over the square of height. This metric classifies individuals into obese classes I, II, III representing stage of obesity respectively, however this measure does not explain the variation within these groups across other factors which have been identified using clustering techniques. This study provides an insight about the factors other than BMI and suggest clinicians to look into them for better decision making and taking the preventive measures to reduce the obesity in the population.

The paper is organized as follows: An overview of a literature survey for outcomes of investigations on obesity in the past and the current ongoing research for its prevention is presented in Related Work. The overall methodology of data collection, clustering methods, and cluster validation indices have been discussed in Material and Methods. Results and Discussion presents the cluster analysis along with the validation process and finalizes the profiling of clustering solution to generate salient characteristics of the obese sample. Potential Recommendations followed by future work in the Conclusion section concludes the paper.

RELATED WORK

The health care sector has been benefitted from machine learning as it has also provided the insight patterns of the characteristics from obese samples. As for Brunei and around the world almost all of the surveys have been using different machine learning techniques to draw the intelligible patterns from the health-related data including the research in the U.S.A, The UK and Brunei Darussalam. The National Health and Nutrition Examination Survey (*NHANES*), administered annually by the National Centre for Health Statistics, is designed to assess the general health and nutritional status of adults and children in the United States of America. These are based on comparisons of data from NHES I (1960-1962), NHANES I (1971-1974), NHANES II (1976-1980) and NHANES Ill (1988-1994). W.H.O consultation on obesity (1999: Geneva, Switzerland) presented NHANES Ill (1988-1994) report and used the obese data having BMI $\geq 30 kg/m^2$ to classify obesity for global comparisons. Different techniques including the least mean square (*LMS*) method of Cole was used for exploring the BMI for ages (James, Bjorntorp, Bray, Carroll, & Chuchalin, 2000). Given to several thousands of individuals, the extent of the NHANES survey is very broad, covering demographic, laboratory and examination information, as well as responses to a fairly comprehensive health questionnaire (Befort, Nazir, & Perri, 2013). NHANES data (1999-2008) was conducted to study the affecting factors (covering demographic, laboratory and examination information, as well as responses to a fairly comprehensive health questionnaire) for NCDs using machine learning techniques including agglomerative hierarchical cluster (*HC*) method (Jun won Lee a, 1, 2013). NHANES 2005-2008 was conducted on a sample of 8815 (*rural, urban*) subjects with 20-75 years of age and measured BMI $\geq 30 kg/m^2$, revealed strong obesity prevalence among the rural-urban residents. Demographic status, diet patterns and physical activity status with BMI measurements were used with clustering probability design using Chi-Square, t-test to conclude the results (Befort et al., 2013). Apart from obesity NHANES has been used in the diagnosis of other non-communicable diseases like a chronic renal failure (*CRD*), cancers, hypertension, DM, etc., (Befort et al., 2013; Delgado, Higuera, Calle-Espinosa, Morán, & Montero, 2017; Green et al., 2015). The National Health Study (*NHS*) from Yorkshire (*2010-2012*) in The UK also used the same kind of approach on 27,806 subjects with 16-85 years of age and measured BMI $\geq 30 kg/m^2$. A Two-Step cluster analysis (*TSC*) was used to define the groups which revealed strong obesity prevalence. Demographic, health and behavioural characteristics with BMI measurements were used to conclude the results (Green et al., 2015). Discussed in first section, NHANSS (*National Health and Nutritional Status Survey*) is a similar kind of survey that is conducted annually to assess the health condition of the population in Brunei Darussalam (MoH, 2014).

MATERIAL AND METHODS

The National Health and Nutritional Status Survey (NHANSS) is conducted annually to access the health and nutritional patterns and characteristics of the population (MoH, 2014). The survey is conducted by the Ministry of Health in Negara Brunei Darussalam which is a cross-sectional survey aimed at the population aged from 5-75 years old with an initial target of 2184 participants from all the districts in Brunei Darussalam. A comprehensive questionnaire was prepared to note down the critical information and the original dataset had 93 features under different labels with 2184 instances under different groups, each addressing the aspects of health and behavior. Demographic information such as gender, age and ethnicity. Face-to-face interviews with parents and/or caregivers (for children) and participants themselves were conducted by trained dietitians/nutritionists and research assistants using a questionnaire booklet with topics such as medical and smoking status, physical activity patterns, history of

raised blood pressure and high cholesterol, health status, body image, food supplements, anthropometric measurements, multiple dietary patterns and bio-chemical measurements on adults and children. The survey procedures and questionnaire were pilot tested prior to training and finalized accordingly for standardized data collection. More details about the sampling techniques used for data collection and preparation of the questionnaire can be found in (MoH, 2014).

DATA SELECTION

In order to study the characteristics of the obese people within the obese classes (I-II-III), the data set was filtered with the number of people having BMI ≥ 30 kg/m^2, out of the total sample of 2184 instances. A subset data set aged 5-75 years was chosen from the original data based on anthropometric measurements and was filtered with labels such as demographics, socio-economic status, medical and smoking status, physical activity patterns, multiple dietary patterns on adults and children. 453 records were filtered with 20.74% percent in respect to the obesity factor on the basis of evidence-based research on obesity (Crawford et al., 1985; W.H.O, 2010). In the obese sample, the data types for the variables were defined as a quantitative and qualitative measure. The level of measurement for quantitative variables was set as numeric while for qualitative variables the level of measurement was set either nominal (for unordered data) or ordinal (for ordered data). Overall, there were 8 numeric variables (quantitative) and the rest of them were either nominal or ordinal (qualitative). The quantitative variables were explored for their collinearity issues with each other at 0.05 threshold significance level which exhibited significant (absolute) correlation coefficients with all the quantitative variables.

TABLE 1. Obese Classification for the Subset Data Set

Obesity Factor	Frequency	Percent	Valid Percent	Cumulative Percent
Obese Class I (30 - 34.9 Kg/m^)	295	65.1	65.1	65.1
Obese Class II (35 - 39.9 Kg/m^)	105	23.2	23.2	88.3
Obese Class III (>=40 Kg/m^)	53	11.7	11.7	100.0
Total	453	100.0	100.0	

The obese sample in Table 1 shows an overview of all the obese classes with frequencies and their percentage in each group of obese classes I, II and III respectively. As depicted, out of the total valid 453 subjects, 295 belonged to obese class I (65.1%) being highest among the classes, 105 being second highest belonged to obese class II (23.2%) and 53 subjects belonged to obese class III (11.7%) being the lowest among all obese classes.

CLUSTERING METHODS

Clustering algorithms can be of two types either hierarchical or non-hierarchical (Larose & Larose, 2014). The choice of a clustering technique depends on the type of data available for a given problem. Since our data set had both numeric and categorical variables, so we chose those clustering techniques which could support the handling of the mixed data types. Upcoming sections will go through the methodology of the distance measure and clustering methods as mentioned below.

DISTANCE MEASURES ~ CLUSTERING MODELS

The clustering techniques use distance measure and it has to be chosen as a function of the data (what they represent) and it also depends on the nature of the data. For instance, agglomerative clustering uses distance measure and places the clusters together as per their closet distance from other clusters. When working with metric (or ordinal data) it uses either Euclidean distance, City-block distance often called Manhattan distance or Chebyshev distance. The choice of the distance metric is very important and should be chosen according to the data types(He, Xu, & Deng, 2005; Jonathan M. Garibaldi, Lai, Malik, Ong, & Wong.). Since the data were mixed types (i.e. numeric and categorical) for this study, all the clustering techniques used Gower's distance as a distance metric. TSC is only available in SPSS platform so the level of measurement for variables and distance metric were defined through its interface.

GOWER'S DISTANCE MEASURE

Gower's Distance was most appropriate for handling the mixed nature of the data (i.e., quantitative and qualitative instances) for cluster analysis because it represents the qualitative data in terms of category matching and quantitative data in terms of geometric distance (He et al., 2005; Jonathan M. Garibaldi et al., 2017). Gower's distance metric calculates the components of the distance between two instances X_i and X_j differently for each variable i.e., categorical instance and continuous instance respectively. For example, consider two instances X_i and X_j both having two variables, denoted by X_{ik} and X_{jk} for $k \in \{1, 2\}$. Assume the first variable is categorical and the second is continuous. For the first variable (the categorical variable), the difference between the values of X_{ik} and X_{jk} is defined as an indicator function as depicted in Equation 2.1 (Gower, 1971):

$$Dijk = \{0 \; if \; X_{ik} = X_{jk} \; 1 \; if \; X_{ik} \neq X_{jk} \tag{2.1}$$

Equation 2.1 shows that if the two cases *i* and *j* are equal then $D_{ijk} = 0$ otherwise $D_{ijk} = 1$. For the second type of variable (the numeric variable), the difference between the values of X_{ik} and X_{jk} is defined in Equation 2.2 as,

$$Dijk = \frac{|X_{ik} - X_{jk}|}{r_k} \tag{2.2}$$

Here, r_k is defined as the range of variable k, max ($x_{.k}$) – min ($x_{.k}$). These two types of variables are the only ones which have been discussed here and since they are the only relevant ones for this thesis though Gower's metric is capable of dealing with other types of variables too (Podani, 1999; Pavoine, et al., 2009). The next step is to combine these D_{ijk} values into Gower's metric (Equation 2.3),

$$Dij = \frac{\sum_{k=1}^{N} w_{ijk} D_{ijk}}{\sum_{K=1}^{N} w_{ijk}} \tag{2.3}$$

where:
- W_{ijk}: the weight for variable k between observations X_i and X_j.
- D_{ijk}: the difference between X_{ik} and X_{jk}.

It is important to note that we use $w_{ijk} = w_k$ when D_{ijk} is defined, effectively assigning one weight per variable. If D_{ijk} is not defined (because there are missing values in the data) then w_{ijk} is equal to 0 (He et al., 2005; Jonathan M. Garibaldi et al., 2017; Yang, 2012). The aforementioned distance measure was used for the clustering solutions as a primary requirement for calculating the distances for all the clustering solutions. For PAM, Gower's distances are calculated between the instances to the cluster average and the nearest instance is chosen as its cluster center. For HC, the distances between data points are calculated using

Gower's distance. For PAM and HC, the distance metric was updated by the help of daisy function which actually takes care of the data normalization and uses Gower's distance as a distance metric available in the R 'cluster' package. The daisy method takes the data matrix as input and the Gower measure is specified as the parameter option.

TWO – STEP CLUSTERING TECHNIQUE

Two-Step clustering method is based on the Balanced Iterative Reducing and Clustering using Hierarchies BIRCH technique (He et al., 2005; Jonathan M. Garibaldi et al., 2017; Zhang, Ramakrishnan, & Livny, 1997). The two-step algorithm performs the clustering in two-stages. During the first stage (pre-clustering), the dense regions of the large dataset are compressed and sub-clusters are formed by using a sequential cluster method. In the second stage, the agglomerative hierarchical clustering method is used to generate the desired number of clusters by using the sub-groups formed in the first stage. The number of clusters (k) can be determined by the algorithm automatically or manually by a user which evaluates these clusters based on the Schwarz's Bayesian's Information Criterion (BIC) and Akaike's Information Criterion (AIC) in an automated process. (Azlan Othman, 2017; M. Hummel, D. Edelmann, 2017). \

HIERARCHICAL CLUSTERING TECHNIQUE

Unlike partition clustering which requires the user to define the value for k (number of clusters) and create homogenous k groups, hierarchical clustering defines clustering using the automated process of distance measuring in the data and defines trees of clusters (Yang, 2012). Hierarchical clustering has two methods (agglomerative and divisive) for groups' formation. Agglomerative clustering algorithm starts with a number of cases that are taken equal to the number of clusters. It merges the clusters, one by one until there is the only cluster that corresponds to the entire data set. The divisive clustering is the opposite of agglomerative clustering. It starts with one cluster which is then processed into two clusters by calculating the distance in the data; these new clusters are divided until it breaks to each case as a cluster (Kavakiotis et al., 2017; Ripley, 2003; Zhang et al., 1997).

PARTITION AROUND MEDOIDS (PAM)

PAM operates on the distance matrix as well as the in-built or the original data matrix which gives it liberty to use the instances of mixed data types i.e., qualitative and quantitative data types. This algorithm actually works like k-means algorithm but with a difference that the latter one uses the original data matrix (Brock, Pihur, Datta, & Datta, 2011). It uses the most central observation rather than the centroid using minimized distances of instances of the respective clusters. Due to the nature of the data set, this clustering solution was one of the choices for the cluster formation using Gower's distance to take care of mixed data types by keeping the structure of the data (Jonathan M. Garibaldi et al., 2017).

KOHONEN NEURAL NETWORKS – SELF ORGANIZING MAPS

Self Organizing Maps are a neural network based clustering technique which is also referred to as Kohonen SOM or Kohonen artificial neural networks (Larose & Larose, 2014; Ripley, 2003). SOM is very useful for clustering and data visualization. Typically, visualizations of SOMs are colorful 2D diagrams of ordered hexagonal nodes and can be applied for supervised or unsupervised learning (Larose & Larose, 2014; Zhang et al., 1997). SOMs do not allow looping so follows the feedforward network and are completely connected to one another in

the next layer as shown in Figure 7. These neurons are standardized so that certain outliers in variables do not influence the smaller values for the output. Since the filtered obese data from MOH did not have any prior information about the data so it was assumed that data set is unsupervised data and accordingly, the SOMs technique has been applied to visualize the data.

CLUSTER VALIDATION

The process of evaluating the quality of clusters is of great importance in machine learning and plays an important role in proposing the best number of clusters for working in the cluster solution. The process is known as cluster validation and has been carried out by the help of cluster validity indices (CVIs) (Guo, Chen, Ye, & Jiang, 2016). It uses the validity indices as a guiding parameter and to cross verify the best number of clusters selected (Ghazzali, 2014). There are few indices that can handle a data set with categorical attributes or both categorical and numeric attributes. There are mainly two types of validity indices measurements i.e. internal and external cluster validation indices (Brock et al., 2011).

TABLE 2. Calculation Parameter for Respective Validity Indices

Sr#	Name of Validity Index	Computation Value
1	Calinski and Harabasz Index (ch)	The Maximum value of the index
2	Davies and Bouldin Index (db)	The Minimum value of the index
3	Hartigan Index	The Maximum difference between hierarchy levels of the index
4	McClain and Rao Index (mcclain)	The Maximum value of the index
5	Krzanowski and Lai Index (kl)	The Maximum value of the index
6	Silhouette Index	The Maximum value of the index
7	Dunn Index	The Maximum value of the index
8	Halkidi et al. 2000 (Sdindex)	The Minimum value of the index
9	Halkidi and Vazirgiannis 2001 (Sdbw)	The Minimum value of the index
10	Duda Index	The Smallest no. of clusters such that index > Critical Value
11	C – Index	that index > critical value
12	Gamma Index	The Maximum value of the index
13	Beale Index	No. of cluster such that critical value >= alpha
14	Hubert Index	No. of cluster represented by the graph knee
15	D – Index	No. of cluster represented by the graph knee

SELECTION CRITERIA FOR BEST NUMBER OF CLUSTERS

In this study, internal indices have been used to evaluate the cluster solutions as there were no data labels (i.e. output) available. Without true cluster labels, estimating the number of clusters (k) in a given data set is a central task in cluster validation and internal indices are preferred e.g. Silhouette Index, Dunn Index, Davis Bouldin Index, etc., (Joe, 2008; Lantz, 2015; Shane Lynn, 2014; Warmbrod, 2001). Depicted in Table 2, fifteen most commonly used internal validity indices have been used which are ranked best in a study by Milligan and Cooper (Yang, 2012). The geometric-like distance was defined using Gower distance and was applied with

numeric indices to explore the data. By utilizing the packages available in R-programing, simultaneous multiple validation measures for the best number of clusters were called in a single function to determine most appropriate optimal number of clusters for a data set such as; ClusterCrit for internal and external indices (Desgraupes, 2013), NbClust (Ghazzali, 2014), clValid (Brock et al., 2011) and clv (Nieweglowski, 2015) for internal indices respectively. For the purpose of CVIs in this research, the values have been selected accordingly with Gower's distance measure as a distance metric for mixed datatypes. The values for the optimal number of clusters was set from 2 to 6 i.e. k=2:6. Table 2 indicates the indices criteria for selecting the best number of clusters.

INTERPRETATION OF CLUSTERS & ANALYSIS

The interpretation of results is the last step in the clustering methodology that is carried out only when the association between the factor variable/s and the found clusters turns out to be statistically significant. It's important to note that if the results are not statistically significant then the process has to be stopped as profiling the non-significant results will be of no use and those results would not be of any importance. A different course of action has to be taken so that the generated patterns show some results being significant. The association has to be checked with the threshold value of alpha which is normally taken as ≤ 0.05 or ≤ 0.01 represented by the Chi-square in the significance tests. For this research, the threshold value of alpha was taken as ≤ 0.05 which is the error rate. The statistically significant results that data depicts would eventually be helpful in knowing the salient characteristics of a specific group of individuals. The significant effect value of Phi and Cramer's V has been checked as presented by Table 3. This has also been as defined by different authors such as (Bartz, 1994), (Hopkins, 1997) and (Cohen, 1988), etc. in (Warmbrod, 2001) while (Hopkins, 1997) convention for the effect size of significance was considered for this research (Warmbrod, 2001).

Table 3. Conventions for Effect Size

Value or r	**Description**
0.9 - 1.0	Nearly perfect, Distinct
0.7 - 0.9	Very Large, Very High
0.5 - 0.7	High, Large Major
0.3 - 0.5	Moderate. Medium
0.1 - 0.3	Low, Small, Minor
0.0 - 0.1	Trivial, Very Small, Insubstantial

RESULTS & DISCUSSION

Since four clustering techniques were applied to the same data set of obese samples, a comparison was made to see the similarities or differences in the results generated by these clustering techniques. The following clustering algorithms were applied to the obese sample for analyzing clusters. Two-Step Clustering (TSC), Agglomerative Hierarchical Clustering HC), Partition Around Medoids (PAM) and Kohonen Self-Organizing Maps (SOM).

CLUSTER EVALUATION CRITERIA

The techniques that were used to evaluate the clustering solutions were internal CVIs to measure the cluster quality, association measures using asymptotic value for significance to measure the association between clustering solutions and obesity classes, Phi and Cramer's V coefficient to measure the effect size between the clustering solutions and obesity and clusters sizing using cluster distribution.

CLUSTER QUALITY MEASURES

Cluster validation indices were used as a guiding parameter to know the optimal number of clusters for a clustering solution to work with. The best number of clusters was proposed as 2 for most of the clustering solutions. Figure 1 represents the best number of clusters proposed by CVIs for all clustering solutions i.e., TSC, HC, PAM, and SOM. It is noticeable that TSC, HC, and PAM proposed 2 as an optimal number of clusters while SOM proposed 3 as optimal number of clusters for the respective solution. On the other hand, Hierarchical and PAM clustering were noted with no difference in projected CVI scores hence HC line is overlapped by the PAM in the figure.

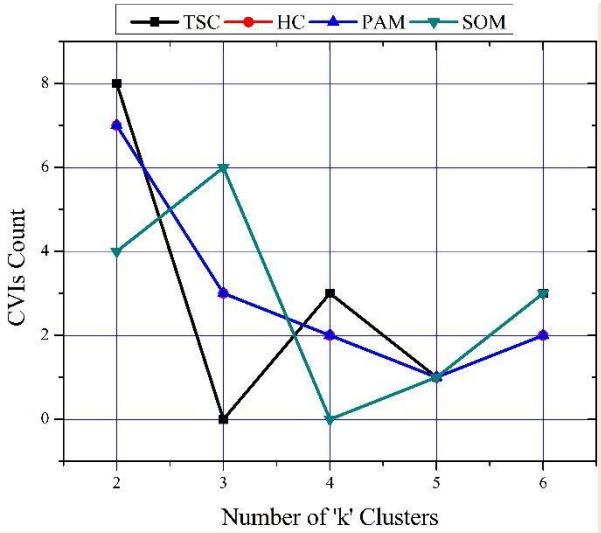


FIGURE 1. Clustering Models ~ CVI Scoring

ASSOCIATION MEASURES

The association of the obesity factor with the clustering solutions as depicted by Figure 2 shows the significance measures of obesity factor with clustering models. It can clearly be seen for TSC that 2-clusters and 3-clusters solution showed significance while there was no significance observed for 4, 5 and 6 clusters solutions. HC showed significance with 2-clusters and 6-clusters solution with no significance observed for 3, 4, and 5 clusters solutions. On the other hand, PAM totally failed to show any significance of the target variable with clustering solutions ranging from 2 to 6 clusters solutions. Significance measure for SOMs could not be measured as its graphical representation was presented.

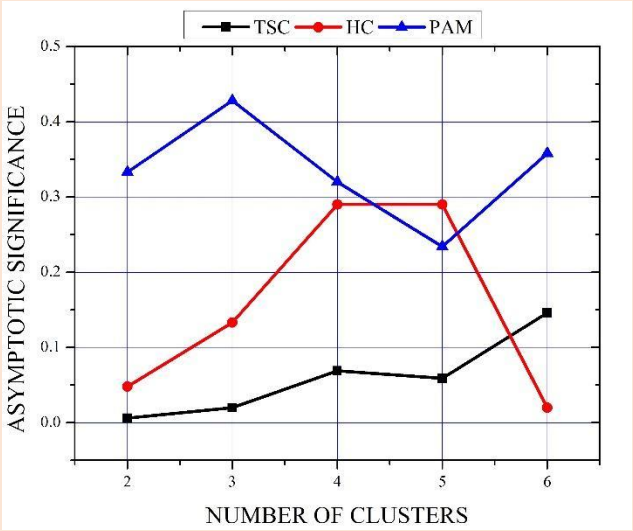


FIGURE 2. Clustering Solution ~ Significance Measures

CLUSTER EVALUATION USING CRAMER'S V COEFFICIENT

Cramer's V coefficient score was calculated using the obesity factor and clustering solutions from respective techniques. A coefficient value near 1 indicates high association while 0 indicates no or less association between the factor subgroups. Table 4 indicates the association measures in terms of Cramer's V and asymptotic values (p-value) between obesity and clustering solutions generated by all the algorithms. The results generated by TSC and HC showed small to the medium effect of significance while PAM was unable to show any significance for the best number of clusters proposed by CVIs and hence the effect shown has been insubstantial. SOM presented 3-clusters solutions and was graphically analysed and discussed.

TABLE 4. Proposed No. of Clusters 'k' ~ Significance Measures

Sr#	Clustering Methods	CVIs b_no_c	CV	Asymp value	Association Measures
1	Two - Step Clustering	2	0.150	0.006	Significant – Small effect
2	Hierarchical Clustering	2	0.116	0.048	Significant –Small effect
3	k - Medioids (PAM)	2	0.07	0.333	Non – Significant – Very Small effect
4	SOMs	3			Graphical Representation

A snapshot in Table 5 has been provided for the association and significance measure of all clustering solutions with respect to the clustering methods ranging from 2 to 6. The table further depicts small to the medium association for TSC and HC (2 and 3 clustering solutions and 2 and 6 clustering solutions respectively) while 2-clusters solution was proposed and discussed for both of the techniques. PAM could not generate any significant results for any cluster solutions but Cramer's V Coefficient depicted small to medium effect for 2 and 6 clusters solution.

TABLE 5. Clustering Solution Comparison ~ Association Scores, Significance Measures

Clust Algo	Two-Step Clustering		Hierarchical Clustering		Partitional Around Medoids	
k	CV	p-value	CV	p-value	CV	p-value

2	0.150	< 0.006	0.116	< 0.048	0.116	> 0.333
3	0.114	< 0.020	0.088	> 0.133	0.088	> 0.428
4	0.114	> 0.069	0.090	> 0.290	0.090	> 0.320
5	0.129	> 0.059	0.103	> 0.290	0.103	> 0.234
6	0.127	> 0.146	0.152	< 0.020	0.152	> 0.358

CLUSTERS DISTRIBUTION

Finally, the clusters sizing was to be checked for comparison among the clustering solutions. Table 6 represents the distribution of the clusters (i.e. the number of data points in each cluster) with respect to the clustering solutions. It depicts almost same pattern for TSC and PAM while the percent of cohort represented by hierarchical clustering was observed a bit different. HC represented almost 68% percent of cases in C1 which was comparatively noted higher than TSC and PAM in C2 otherwise the distribution was noted in a similar pattern.

TABLE 6. Cluster Distribution ~ Cluster Solutions

Sr#	Clustering Solution	C1	C2	Total	% C1	% C2
1	Two - Step Clustering	191	262	453	42.16%	57.84%
2	Hierarchical Clustering	306	147	453	67.55%	32.45%
3	k - Medoids Clustering	226	227	453	49.89%	50.11%
4	Self-Organizing Maps	Graphical Representation				

CLUSTERING SOLUTIONS

Since four clustering techniques were used for NHANSS obese sample analysis so the most prevalent conditions generated within the subgroups have been discussed below. As aforementioned PAM did not show any association of target variable (obesity) with that of the clustering solution so it has not been further discussed while SOMs was analysed visually leaving behind TSC and HC for further analysis on most prevalent conditions within generated subgroups. Figures 3 and 4 present the distribution of the cases with respect to the class wise percentage representation within respective clusters C1 and C2. We can see the distribution of the cluster that shows class I, class II and class III respectively alongside percent of the cases clustered in respective classes. Considering TSC cluster 1, around 48% of case are clustered in class I while class II and III cases have been relatively low in percentage i.e. 34% and 28% respectively (i.e., fewer cases were clustered in C1, TSC model) while in HC model, the majority of class II and class III cases are grouped together in cluster 1 with around 77% and 68% respectively while class I with 64% percent being lowest than the other classes (i.e., more cases were clustered in C1, HC model).

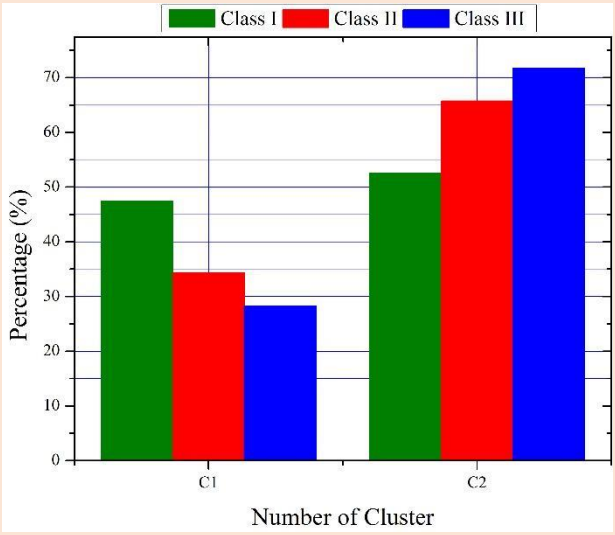


FIGURE 3. Obesity Classes ~ Two-Step Clustering Model

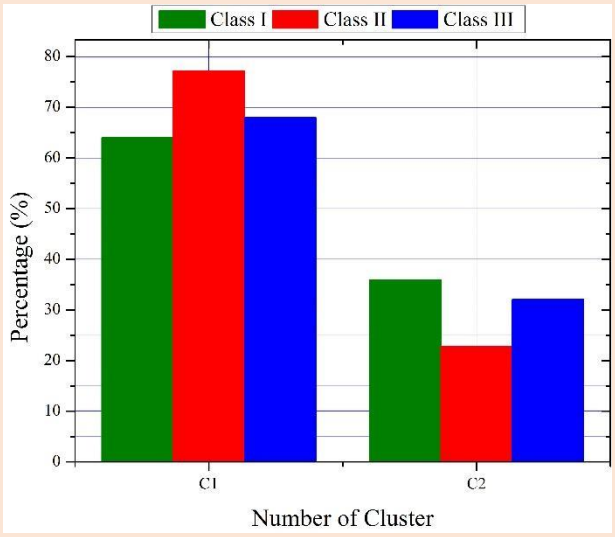


FIGURE 4. Obesity Classes ~ Hierarchical Clustering Model

On the other hand, if we have a look at TSC cluster 2 it shows around 53% of class I cases, 66% of class II cases and 72% of class III cases clustered respectively (i.e., more cases were clustered in C2, TSC model) while in HC model, the majority of class I and class III cases grouped together with around 36% and 32% respectively while class II with 23% percent being lowest than the other classes (i.e., fewer cases were clustered in C2, HC model).This showed that cluster 1 for TSC was less densely populated than cluster 2 in TSC model and cluster 1 in HC model representing a less obese group of individuals while cluster 2 had more percentage of more obese individuals compared to cluster 1 in TSC and cluster 2 in HC model representing the obese group of individuals. Apart from the cluster analysis, important predictors together with numeric variables had played an important role for both the clustering models (TSC and HC) while the association of all of these variables was noted statistically significant. The profiling of these quantitative variables further helped in categorizing the distinct characteristics within the subgroups as mentioned below. In order to analyze the behavior of quantitative variables across each cluster along with the results of an analysis of variance for both finally selected techniques (i.e. TSC and HC), results have been generated for their means, standard deviation, standard error ($standard\ deviation/\sqrt{cluster\ count}$), 95% confidence

interval of mean value, minimum and maximum value of each variable. The labels of these quantitative variables are mentioned in Table 7 for better understanding.

TABLE 7. Cluster Distribution ~ Cluster Solutions

Sr #	Variable	Label
1	ageyears	Age of the respondents
2	RstM49b	How much time do you usually spend sitting, resting on a typical day?
3	TvM50b	How much time do you usually spend watching television on a typical day?
4	FrtDy77	In a typical week, how many days do you eat fruits?
5	FrtSv78	How many servings of fruits do you eat on one of those days?
6	VgDy79	In a typical week, how many days do you eat vegetables?
7	VgSv80	How many servings of vegetables do you eat on one of those days?
8	OutFd81	On average, how many meals per week do you eat that were not prepared at home? By meals it means breakfast, lunch or dinner.

DEMOGRAPHIC CHARACTERISTICS

The average age and marital status were important predictors revealed by TSC and HC while residential status was only revealed by HC as important predictor as far as demographic characteristics were concerned. For TSC the average age of cluster 1 was 45 years showing that most of the individuals belonged to this group were middle-aged and married while the average age of HC cluster 1 was 33 years, means it had most of the individuals who were young aged and single. Comparatively, the average age of TSC cluster 2 was 29 years, consisting of young aged, single individuals while the average age of HC cluster 2 was 43 years means they were middle-aged, married individuals. Residential status showed association with HC model only which revealed that most of the individuals in cluster 1 were Brunei Citizen while in cluster 2, most of the individuals were permeant residents which depict that the sample size was tested on individuals who were either Brunei citizens or permanent resident. The mean values for average age in TSC and HC models can further be verified by looking at Tables 8 and 10 (column 1, line 2) together with Figures 5 (A) and 6 (A), respectively. A closer look at the cluster solutions indicate that here were mainly two subgroups: one belonged to an active middle-aged group of individuals who lived in rural areas while the other belonged to a young age group of individuals who lived in urban areas.

PHYSICAL INACTIVITY CHARACTERISTICS

The next two quantitative variables define the physical activity patterns of the respective groups in the respective clustering solutions (TSC & HC). For TSC, the question about resting/reclining, Tables 8 and 10 depict that the individuals in TSC cluster 1 on average spent 213 minutes (3.55 hours ≈ 4 hours a day) while individuals in HC cluster 1 spent 273 minutes (4.55 hours ~ 5 hours a day) in sitting, resting or reclining while individuals in TSC cluster 2 spent on average 294 minutes (5 hours a day), comparatively more time than TSC & HC cluster 1 individuals. On the other hand individuals in HC cluster 2 spent 233 minutes (3.8 hours ~ 4 hours a day) comparatively lesser time than HC cluster 1 individuals so it was concluded that individuals in TSC cluster 1 was the group that spent less time sitting, resting or reclining comparatively to TSC cluster 2 while the individuals in HC cluster 2 was the group that spent less time sitting, resting or reclining on a typical day than HC cluster 1. Figures 5 (C) and 6 (C) verifies the means of the respective variables in the boxplots.

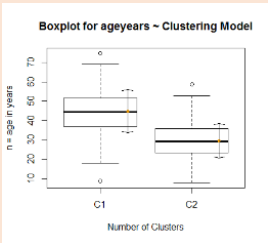
The same trend was observed for the question about time spent to watch TV. It was found that most of the individuals in cluster 1 spent on average 102 minutes (1.7 hours ≈ 2 hours a day) and the individuals in HC cluster 1 spent 122 minutes (2.03 hours a day) in watching TV on a

typical day while individuals in TSC cluster 2 spent on average 123 minutes (2.05 hours a day), comparatively more time than TSC cluster 1 but almost same time as HC cluster 1 while individuals in HC cluster 2 spent 98 minutes (1.63 hours a day) comparatively lesser time than HC cluster 1 so it was concluded that individuals in TSC cluster 1 was the group that spent less time watching TV on a typical day than TSC cluster 2 but more time than HC cluster 1 while HC cluster 2 was the group that spent lesser time watching TV on a typical day than HC cluster 1. The mean values have been depicted by Tables 8 and 10 together with Figures 5 (B) and 6 (B) respectively.

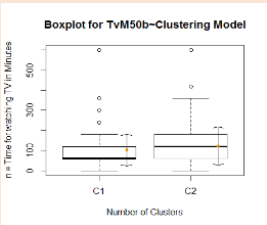
SHORT FOOD FREQUENCY CHARACTERISTICS

Rest of the remaining 5 quantitative variables inform us the short food frequency habits of the sample in the clustering solutions (TSC & HC). As depicted in Tables 8 and 10 and shown in Figures 5 (D - H) and Figures 6 (D - H), the variables are discussed in sequential order. The variables, “In a typical week, how many days do you eat fruits?” and "How many servings of fruits do you eat on one of those days?" showed that the individuals in TSC cluster 1 ate fruits 3 days in a week and almost 2 servings of fruits on one of those days while individuals in TSC cluster 2 ate fruits 2 days in a week and 1 serving of fruits on one of those days respectively, comparatively less than TSC cluster 1 individuals. The pattern was observed vice versa in HC solution where the individuals in HC cluster 1 ate fruits 2 days in a week and almost 1 serving of fruits on one of those days while individuals in HC cluster 2 ate fruits, 4 days in a week and 2 servings on one of those days respectively, comparatively more than HC cluster 1 individuals so it was concluded that TSC cluster 1 was a group of individuals whose fruit intake was more than the TSC cluster 2 group of individuals in a week. (Figures 5 D & E represents the details of means for TSC cluster 1 and cluster 2 respectively) while HC cluster 1 was the group of individuals whose fruit intake was less than the individuals of HC cluster 2 in a week. (Figures 6 D & E represents the details of means for HC cluster 1 and cluster 2 respectively).

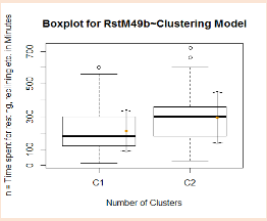
Quantitative Variables ~ TSC Solution



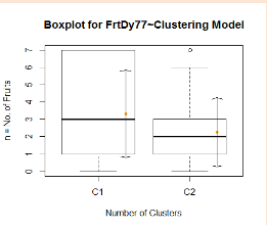
A. Age in years ~ TSC Model



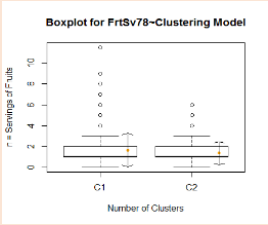
B. How much time do you usually spend watching television on a typical day?



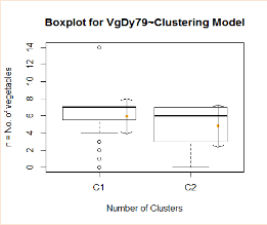
C. How much time do you usually spend sitting, resting or reclining on a typical day?



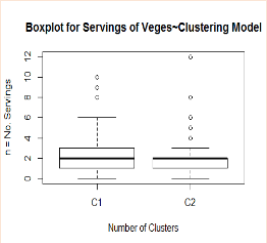
D. In a typical week, how many days do you eat fruits?



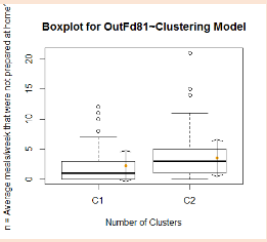
E. How many servings of fruits do you eat on one of those days?



F. In a typical week, how many days do you eat vegetables?



G. How many servings of vegetables do you eat on one of those days?



H. On average, meals per week that were not prepared at home?

FIGURE 5. Quantitative Variables ~ TSC Solution

For vegetable intake and servings of vegetables in a week, almost the same trend was observed. The variables “In a typical week, how many days do you eat vegetables?” and "How many servings of vegetables do you eat on one of those days? (1 serving = half cup of cooked vegetables / 1 cup of salad vegetable)" as shown in Figures 5 (F, G). TSC cluster 1 clearly leads cluster 2 who took vegetables 6 days in a week and almost 3 servings of vegetables on one of those days while individuals in TSC cluster 2 ate vegetable 5 days in a week and 2 servings of vegetables on one of those days respectively which was found comparatively less than TSC cluster 1 individuals. On the other hand as shown in Figures 6 (F, G), HC cluster 2 clearly dominated cluster 1 who took vegetables for more than 6 days in a week and almost 3 servings of vegetables on one of those days while individuals in HC cluster 1 ate vegetable 5 days in a week and 2 servings of vegetables on one of those days respectively, comparatively less than cluster 2 individuals.

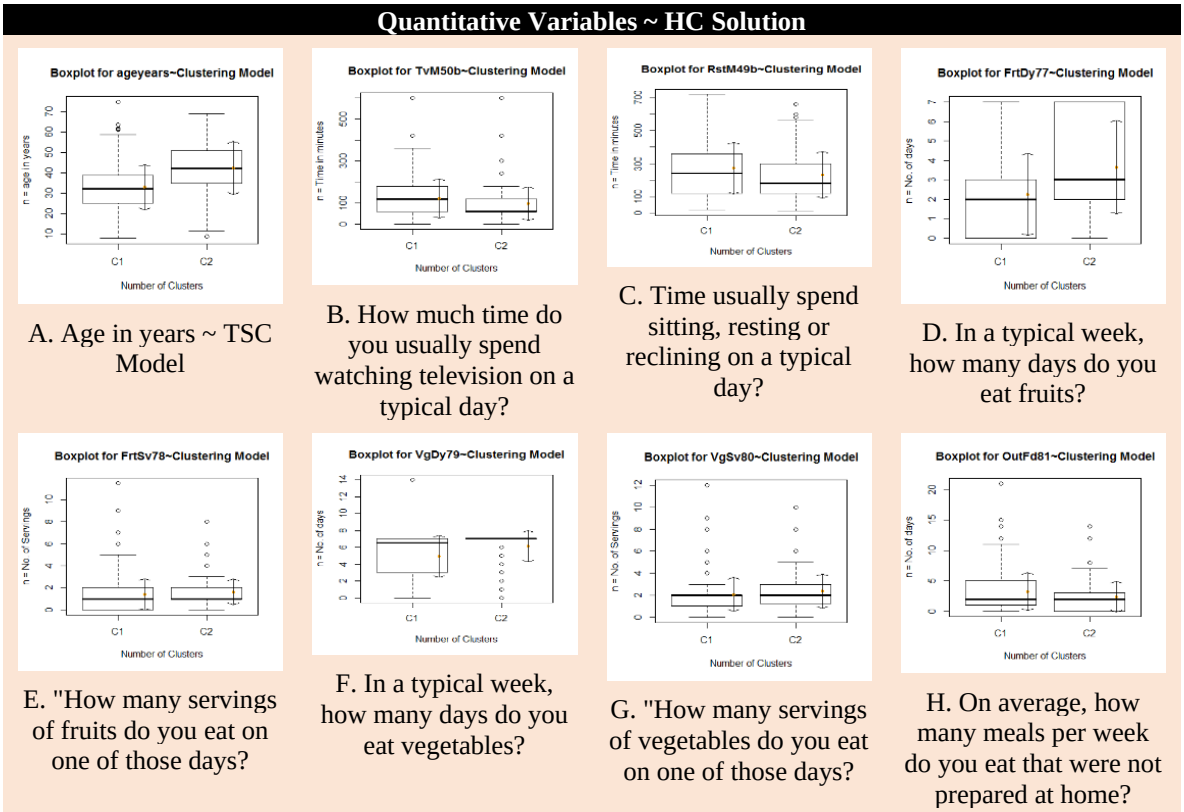


FIGURE 6. Quantitative Variables ~ HC Solution

For average meals (breakfast, lunch or dinner) prepared in home in a day, the trend of number of meals prepared at home was more in TSC cluster 2 than cluster 1. Individuals in TSC cluster 1 prepared 2 meals in home means they were taking 2 home-cooked meals at home in a day while TSC cluster 2 prepared 4 meals in the home means they were taking 4 home-cooked meals at home in a day. On the other hand, the trend of a number of meals prepared at home was more in HC cluster 1 than HC cluster 2. The value of means is depicted by the help of boxplot in Figures 5 (H) and 6 (H) respectively.

TABLE 8. Descriptives of Continuous Variables in the Data Set ~ TSC Solution

Variables	Cluster No	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min	Max
						Lower Bound	Upper Bound		
ageyears	1	191	44.8112	11.22273	.81205	43.2094	46.4130	8.82	74.71
	2	262	29.5282	8.94125	.55239	28.4405	30.6159	7.95	58.84
	Total	453	35.9720	12.49766	.58719	34.8181	37.1260	7.95	74.71
RstM49b	1	191	213.0848	127.29168	9.21050	194.9168	231.2528	15.00	600.00
	2	262	293.5878	158.55402	9.79550	274.2995	312.8761	30.00	720.00
	Total	453	259.6450	151.36469	7.11173	245.6689	273.6212	15.00	720.00
TvM50b	1	191	101.9476	79.01244	5.71714	90.6704	113.2249	.00	600.00
	2	262	122.7176	96.32069	5.95071	111.0000	134.4351	.00	600.00
	Total	453	113.9603	89.92735	4.22516	105.6569	122.2637	.00	600.00
FrtDy77	1	191	3.33	2.534	.183	2.97	3.69	0	7
	2	262	2.26	2.002	.124	2.01	2.50	0	7
	Total	453	2.71	2.301	.108	2.50	2.92	0	7
FrtSv78	1	191	1.64	1.595	.115	1.41	1.87	0	12
	2	262	1.35	1.111	.069	1.21	1.48	0	6
	Total	453	1.47	1.343	.063	1.35	1.59	0	12
VgDy79	1	191	5.97	2.052	.148	5.68	6.26	0	14
	2	262	4.81	2.443	.151	4.52	5.11	0	7
	Total	453	5.30	2.354	.111	5.08	5.52	0	14
VgSv80	1	191	2.42	1.642	.119	2.19	2.66	0	10
	2	262	1.99	1.433	.089	1.82	2.17	0	12
	Total	453	2.17	1.538	.072	2.03	2.32	0	12
OutFd81	1	191	2.19	2.572	.186	1.82	2.56	0	12
	2	262	3.53	3.062	.189	3.16	3.90	0	21
	Total	453	2.96	2.939	.138	2.69	3.24	0	21

TABLE 9. Quantitative Variables ~ ANOVA (Analysis of Variance) ~ TSC Solution

Variables		Sum of Squares	Df	Mean Square	F	Sig.
ageyears	Between Groups	25802.238	1	25802.238	259.772	.000
	Within Groups	44796.316	451	99.327		
	Total	70598.555	452			
RstM49b	Between Groups	715913.014	1	715913.014	33.494	.000
	Within Groups	9639980.227	451	21374.679		
	Total	10355893.241	452			
TvM50b	Between Groups	47654.709	1	47654.709	5.957	.015
	Within Groups	3607636.576	451	7999.194		
	Total	3655291.285	452			
FrtDy77	Between Groups	127.450	1	127.450	25.365	.000
	Within Groups	2266.086	451	5.025		
	Total	2393.536	452			
FrtSv78	Between Groups	9.505	1	9.505	5.321	.022
	Within Groups	805.563	451	1.786		
	Total	815.067	452			
VgDy79	Between Groups	147.523	1	147.523	28.220	.000
	Within Groups	2357.647	451	5.228		
	Total	2505.170	452			
VgSv80	Between Groups	20.589	1	20.589	8.859	.003
	Within Groups	1048.134	451	2.324		
	Total	1068.723	452			

	Between Groups	198.964	1	198.964	24.223	.000
OutFd81	Within Groups	3704.470	451	8.214		
	Total	3903.435	452			

TABLE 10. Descriptives of Continuous Variables ~ HC Solution

Variables	Cluster No	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min	Max
						Lower Bound	Upper Bound		
	1	306	32.8986	11.06677	.63265	31.6537	34.1435	7.95	74.71
ageyears	2	147	42.3698	12.91086	1.06487	40.2652	44.4743	8.82	69.01
	Total	453	35.9720	12.49766	.58719	34.8181	37.1260	7.95	74.71
	1	306	272.6961	154.98229	8.85975	255.2621	290.1300	20.00	720.00
RstM49b	2	147	232.4776	140.18884	11.56258	209.6259	255.3292	15.00	660.00
	Total	453	259.6450	151.36469	7.11173	245.6689	273.6212	15.00	720.00
	1	306	121.8660	93.56515	5.34876	111.3409	132.3912	.00	600.00
TvM50b	2	147	97.5034	79.65038	6.56945	84.5199	110.4869	.00	600.00
	Total	453	113.9603	89.92735	4.22516	105.6569	122.2637	.00	600.00
	1	306	2.25	2.110	.121	2.02	2.49	0	7
FrtDy77	2	147	3.65	2.400	.198	3.26	4.04	0	7
	Total	453	2.71	2.301	.108	2.50	2.92	0	7
	1	306	1.40	1.416	.081	1.24	1.55	0	12
FrtSv78	2	147	1.62	1.165	.096	1.43	1.81	0	8
	Total	453	1.47	1.343	.063	1.35	1.59	0	12
	1	306	4.91	2.475	.141	4.63	5.19	0	14
VgDy79	2	147	6.11	1.840	.152	5.81	6.41	0	7
	Total	453	5.30	2.354	.111	5.08	5.52	0	14
	1	306	2.08	1.526	.087	1.91	2.25	0	12
VgSv80	2	147	2.37	1.548	.128	2.12	2.63	0	10
	Total	453	2.17	1.538	.072	2.03	2.32	0	12
	1	306	3.24	3.091	.177	2.89	3.59	0	21
OutFd81	2	147	2.39	2.509	.207	1.99	2.80	0	14
	Total	453	2.96	2.939	.138	2.69	3.24	0	21

TABLE 11. Quantitative Variables ~ ANOVA (Analysis of Variance) ~ HC Solution

Variables		Sum of Squares	df	Mean Square	F	Sig.
	Between Groups	8907.339	1	8907.339	65.118	.000
ageyears	Within Groups	61691.216	451	136.788		
	Total	70598.555	452			
	Between Groups	160617.510	1	160617.510	7.105	.008
RstM49b	Within Groups	10195275.731	451	22605.933		
	Total	10355893.241	452			
	Between Groups	58937.030	1	58937.030	7.391	.007
TvM50b	Within Groups	3596354.255	451	7974.178		
	Total	3655291.285	452			

	Between Groups	194.113	1	194.113	39.804	.000
FrtDy77	Within Groups	2199.424	451	4.877		
	Total	2393.536	452			
	Between Groups	5.118	1	5.118	2.850	.092
FrtSv78	Within Groups	809.950	451	1.796		
	Total	815.067	452			
	Between Groups	142.294	1	142.294	27.159	.000
VgDy79	Within Groups	2362.876	451	5.239		
	Total	2505.170	452			
	Between Groups	8.684	1	8.684	3.694	.055
VgSv80	Within Groups	1060.039	451	2.350		
	Total	1068.723	452			
	Between Groups	70.734	1	70.734	8.323	.004
OutFd81	Within Groups	3832.701	451	8.498		
	Total	3903.435	452			

Based on analysis of Table 9, it was concluded that all the quantitative variables in TSC cluster solution have been found significantly different between the groups and within the groups representing that the values of the variables are significant between the groups (TSC cluster 1 and 2). For instance, looking at variable age the (1^{st} & 7^{th} column, 2^{nd} line) the results depict significant results between the groups which means that age has an influence on the individual being obese similarly the time spent in watching TV and the time spent in resting/reclining also had an obesity impact. All the other numeric variables from short food frequency were also found statistically significant which shows that all of these characteristics had a significant impact on obesity. To conclude Table 9 and 11, all the quantitative variables in TSC and HC cluster solution were found significantly different between the groups and within the groups. It showed that the values of the variables are significant between the subgroups (HC cluster 1 and 2). Considering Table 8 & 10, looking at the (1^{st} & 7^{th} column, 2^{nd} line) variable age, the results depicted significant results between the groups which means that age had an influence on the individuals being obese similarly the time spent in watching TV and the time spent in resting/reclining together with all the other numeric variables also had an obesity impact.

SOMS SOLUTION WITH HIERARCHICAL CLUSTERING ANALYSIS

Since the CVIs scoring was carried out on SOM normalized data for the optimal number of clusters, the proposed optimal number of clusters was 3-clusters solution. Once the training and visualization process was completed, hierarchical clsutering was applied to the SOMs model (Delgado et al., 2017; Unglert, Radi, & Jellinek, 2016). Hierarchical clustering on the SOM was the choice for carrying out the clustering in order to separate the groups of instances that were similar in metrics. This would separate the groups with similar metrics and manual visualization of clusters would be more evident. Hierarchical cluster method was applied by specifying 3-clusters solution, it was also used in other clustering methods like TSC and HC in this study. As depicted in Figure 7 (A), three clusters were generated highlighted by red, orange and yellow colors respectively, with the visuals of data points so that the distribution could be seen clearly. It is noticeable that nodes with more and fewer data points have converged in one cluster and others in others representing the similar metrics (characteristics) by the respective subgroups. The visualization of clusters have been provided in Figure 7 (B).

EVALUATION FACTOR VISUALIZATION ~ SOMS SOLUTION

The "heatmap" are generated to visualize the distribution of evaluation factor (obesity) across the map but it could not be generated because of high dimensional SOM > 7 variables (i.e., heatmaps are not suitable for SOM model having more than 7 variables and since we had 31 variables so it was not possible to visualize obesity factor in all dimensions of one SOM diagram) (Shane Lynn, 2014). SOM, on the other hand, provided 3 clusters. Figures 7 (A – B) seemingly depicts the cluster formation on a good distribution of the data points.

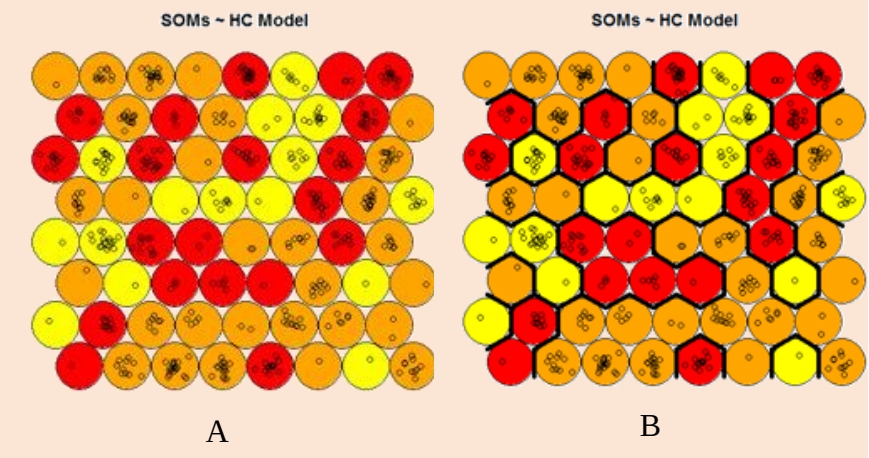


FIGURE 7. Training Data Set ~ Clustering Solution

The distribution with respect to the similar group of instances in respective clusters provided an insight of three subgroups of obesity sharing similar characteristics in respective subgroups. The obesity factor could have been studied quite well with "heatmap" distribution on overall SOM maps but since it was not possible as aforementioned hence the generated subgroups on maps were analyzed visually.

POTENTIAL RECOMMENDATIONS TO CLINICIANS

This research is of clinical importance and the salient features have been reported. The proposed approach reveals the obesity sub-groups which may help to investigate the importance of affecting factors discussed in detail such as demographic status, socio-economic status, anthropometric measurements and behavioral characteristics such as short food frequency and physical activity from the clinical point of view. Clustering techniques used in this research especially Two-Step clustering, Agglomerative hierarchical clustering and SOMs revealed the obesity sub-groups based on affecting factors as depicted in Table 12. These clustering techniques uses hierarchical cluster method which have proved to be one of the optimized methods for cluster analysis as it creates homogenous groups to identify the affecting factor clusters and patterns. HC provides scalable clustering analysis and is designed to handle the big datasets with mixed data instances using relevant distance measure. The analysis of the results provides an insight of HC performance which would be helpful in applying to large data sets to identify and interpret the affecting factors efficiently or it can be a difficult otherwise. That is the reason, it was used in other clustering methods like TSC, HC and SOMs in this study. The trends in the obese subgroups for recreational activities and dietary patterns in middle-aged group were observed totally vice-versa to the young aged group which depicts that the recreational activities together with health dietary intake are two important lifestyle factors of great importance in daily routine life to stay healthy.

CONCLUSION

The clustering analysis was successfully performed on the health and nutritional status survey data for identifying the obese groups from the given data without depending on the standard obesity classification while provided a unique method for investigating the co-occurrences of obese conditions. Working with large data sets simple interpretation of numbers do not presents the conditions or the classes while interpretation is difficult. The clustering techniques helped to uncover the hidden underlying patterns in the data. There were mainly two subgroups identified, one belonged to an active middle-aged group of individuals who lived in rural areas while the other belonged to a young age group of individuals who lived in urban areas of Brunei Darussalam. The first one was reported with controlled diet while the latter one was reported with a sedentary lifestyle and unhealthy diet patterns. This research is of clinical importance and the salient features have been reported but further investigation from a medical perspective is required. The proposed approach reveals the sub-groups which helped to investigate the importance of the physical activities and short food frequency from the clinical point of view. Overall, the combination of clinical knowledge with data-hidden information, as well as the evaluation of subclasses revealed by the data structure could lead to very interesting developments. Two-Step clustering and hierarchical clustering outperformed among the four techniques applied and identified the subgroups based on the underlying hidden patterns in the data while PAM and SOMs could not generate the results for proper identification of the obese subgroups. The future work shall be carried forward in the same direction with the inclusion of improved algorithms such as BIRCH, CATREG or SQUEEZER algorithm can be introduced to interpret the results by comparing with already generated results alongside improvements in the adoption of CVIs to determine the optimal cluster sizes. Another challenging issue was to deal with the algorithmic development framework which is the exploration of the relevant distance metric. Despite the ability of this approach to include and preserve the property of Gower's distance, further investigation of the appropriate merging criteria is needed in order to validate the form of the final results.

TABLE 12. Description of sub-groups identified within obese population using TSC & HC Model

k	Anchoring Conditions	(N)	% of Cohort	Most Prevalent conditions in Clusters
1	Active elderly aged individuals with controlled diet	TSC 191	42.16%	**Demographic Status** - Brunei Darussalam citizens or permanent residents. - Elderly 43-45 years aged less obese group (class – I) - 82.7 were married. **Multiple Dietary Patterns** - 50.8% Rarely ate Chicken Tail / Wings / Skin. - 62.3% Rarely snacks on crisps / keropok.

				- 53.4% Rarely eat Nasi Katok. - All days of the week this group mostly ate vegetables. - All days of the week this group mostly ate fruits. - 53.5% often ate Malay Kuih such as bingka, kusui, seri muka, cucur, cakoi, karipap, popia, kelupis, pie and others? **Physical Activity Patterns** - 91.1% Did not vigorous intensity sports fitness or recreational (leisure) activities. - 62.3% Do moderate intensity sports fitness or recreational (leisure) activities. **Demographic Status** - Brunei Darussalam citizens or permanent residents. - Young 29-33 years aged more obese group (class – II & III) - 50% were Single
2		HC 147	32.45%	
2	Young aged individuals with sedentary life style and unhealthy diet patterns	TSC 262	57.84%	**Multiple Dietary Patterns** - 54.6% mostly ate Chicken Tail / Wings / Skin. - 61.5% mostly ate fast food such as Fried Chicken / Pizza / Chips / Burger / Sausage / Nugget. - 51.5% Rarely eat Nasi Katok. - 63.4% mostly eat instant noodles. - 91.6% of individuals were not told by a doctor or health worker that you have high blood cholesterol? - 55.3% drank fizzy / carbonated drinks / cordials / syrups / sports drink 2 times per week.
1		HC 306	67.55%	- 64% Did not take vegetables in a week in a typical week. **Physical Activity Patterns** - 5 hours were spent by this group usually sitting, resting or reclining on a typical day. - 76.3% Did not vigorous intensity sports fitness or recreational (leisure) activities.

ACKNOWLEDGMENTS

The authors would like to express sincere appreciation to the technical assistance and support from the Department of Economic Planning and Development Brunei Darussalam together with Ministry of Health Brunei Darussalam, and participation from the survey respondents.

CONFLICT OF INTEREST

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

ASEAN Secretariat. (2014). *Association of Southeast Asian Nations, Annual Report, 2013-2014*. (ASEAN Secretariat, Ed.). Jakarta, Indonesia: JAKARATA, ASEAN Secretariat. Retrieved from https://asean.org/storage/2014/07/7.-July-2014-2013-2014-ASEAN-Annual-Report.pdf

Azlan Othman. (2017). Obesity rates in Brunei highest in region | Borneo Bulletin Online. Retrieved April 21, 2019, from https://borneobulletin.com.bn/obesity-rates-in-brunei-highest-in-region/

Befort, C. A., Nazir, N., & Perri, M. G. (2013). Prevalence of Obesity Among Adults From Rural and Urban Areas of the United States: Findings From NHANES (2005–2008). *J Rural Health*, *28*(4), 392–397. https://doi.org/10.1111/j.1748-0361.2012.00411.x.Prevalence

Brock, G., Pihur, V., Datta, S., & Datta, S. (2011). clValid , an R package for cluster validation.

Journal of Statistical Software, (Oct 2011), 1–32.

Crawford, J., Gower, J., Lingoes, J., Rhee, W., Rohlf, F. J., & Sarle, W. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–160.

Delgado, S., Higuera, C., Calle-Espinosa, J., Morán, F., & Montero, F. (2017). A SOM prototype-based cluster analysis methodology. *Expert Systems with Applications*, *88*, 14–28. https://doi.org/10.1016/j.eswa.2017.06.022

Desgraupes, B. (2013). Clustering Indices (pp. 1–33).

Gatta-Cherifi, B. (2016). Obésités : quoi de neuf en 2016 ? Obesities : What ' s new in 2016 ? *Les Must de l'endocrinologie 2016 Sponsor: Publication of This Supplement Was Made Possible with Support from Ipsen-Pharma*, *77*, S29–S35. https://doi.org/10.1016/S0003-4266(17)30075-6

Ghazzali, N. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, *61*(6), http://www.jstatsoft.org/. Retrieved from http://cran.r-project.org/package=NbClust

Green, M. A., Strong, M., Razak, F., Subramanian, S. V., Relton, C., & Bissell, P. (2015). Who are the obese ? A cluster analysis exploring subgroups of the obese. *Journal of Public Health Advance Access Published*, 1–7. https://doi.org/10.1093/pubmed/fdv040

Guo, G., Chen, L., Ye, Y., & Jiang, Q. (2016). Cluster Validation Method for Determining the Number of Clusters in Categorical Sequences. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13. https://doi.org/10.1109/TNNLS.2016.2608354

He, Z., Xu, X., & Deng, S. (2005). Scalable algorithms for clustering large datasets with mixed type attributes. *International Journal of Intelligent Systems*, *20*(10), 1077–1089. https://doi.org/10.1002/int.20108

James, W. P. T., Bjorntorp, P., Bray, G. A., Carroll, K. K., & Chuchalin, A. (2000). *Obesity : preventing and managing the global epidemic : report of a WHO consultation.* Geneva - Switzerland. https://doi.org/ISBN 92 4 120894 5

Joe, H. (2008). Joe , Harry . Multivariate Models and Dependence Concepts . (1997) Outline Multivariate EVD Theory. In *Chapman & Hall*. https://doi.org/0-412-07331-5

Jonathan M. Garibaldi, Lai, D. T. C., Malik, O. A., Ong, S. K., & Wong., J. (2017). Pattern Analysis and Applications A Cluster Analysis of population based cancer registry in Brunei Darussalam : An exploratory study, 1–15.

Jun won Lee a, 1, C. G.-C. b. (2013). Results on mining NHANES data : A case study in evidence-based medicine. *Computers in Biology and Medicine*, *43*(5), 493–503. https://doi.org/10.1016/j.compbiomed.2013.02.018

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, *15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

Lantz, B. (2015). *Machine Learning with R*. (A. Lobo & S. Priya, Eds.) (Second edi). Livery Place 35 Livery Street Birmingham B3 2PB, UK. Retrieved from www.packtpub.com

Larose, D., & Larose, C. (2014). *Discovering knowledge in data, An Introduction to Data Mining* (2nd Editio). Published by John Wiley & Sons, Inc., Hoboken, New Jersey. Retrieved from www.wiley.com.

M. Hummel, D. Edelmann, A. K.-S. (2017). *Clustering and Visualization of Mixed-Type Data* (No. 2.1). Germany: Repository CRAN. Retrieved from m.hummel@dkfz.de

MoH, B. (2014). *The Report, The 2nd National Health and Nutritional Status Survey (NHNANSS) 2014*. Bandar Seri Begawan: Ministry of Health, Common Wealth Drive, Brunei Darussalam.

MOH Brunei. (2018). *BruMAP NCD Brunei 2013-2018*. Bandar Seri Begawan, Brunei Darussalam. Retrieved from http://www.moh.gov.bn/SiteCollectionDocuments/Downloads/downloads/BRUMAPBOOK.pdf

Nieweglowski, L. (2015). Cluster Validation Techniques. *R Repository CRAN*, (2013-11-11 13:44:41), 1–36.

Ong, S. K., Teck, D., Lai, C., Yun, J., Wong, Y., Si-ramlee, K. A., Chb, M. B. (2017). Cross-sectional STEPwise Approach to Surveillance (STEPS) Population Survey of Noncommunicable Diseases (NCDs) and Risk Factors in Brunei Darussalam 2016. *Asia Pacific Journal of Public Health*, *29*(8), 635–648. https://doi.org/10.1177/1010539517738072

Ripley, B. D. (2003). Statistical Data Mining. *Springer-Verlag*, *New York*, (April), 7–40. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4/

Shane Lynn. (2014). Self-Organising Maps for Customer Segmentation using R | R-bloggers. Retrieved April 22, 2019, from https://www.r-bloggers.com/self-organising-maps-for-customer-segmentation-using-r/

Siddiqi, K. (2010). Non-communicable diseases. *Public Health: An Action Guide to Improving Health*. https://doi.org/10.1093/acprof:oso/9780199238934.003.15

Unglert, K., Radi, V., & Jellinek, A. M. (2016). Principal component analysis vs . self-organizing maps combined with hierarchical clustering for pattern recognition in volcano seismic spectra. *Journal of Volcanology and Geothermal Research*, *320*, 58–74. https://doi.org/10.1016/j.jvolgeores.2016.04.014

W.H.O. (2010). Obesity and overweight. https://doi.org/10.3233/978-1-61499-438-1-372

Warmbrod, J. R. (2001). Calculating, Interpreting, and Reporting Estimates of "Effect Size" (Magnitude of an Effect or the Strength of a Relationship). *National Agricultural Education Research Conference*, 1–22.

William Chee Fui CHONG, CHONG, V. H., & PANDE, K. (2013). Brunei International Medical Journal (BIMJ) Official Publication of the Ministry of Health, Brunei Darussalam. *Brunei Int Med J.*, *9*(November).

Yang, R. (2012). *A Hierarchical Clustering and Validity Index for Mixed Data*. Iowa State University. Retrieved from http://lib.dr.iastate.edu/etd/12534%0AThis

Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH: A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery*, *35*(4), 141–182.

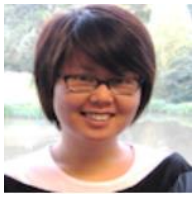
USMAN KHALIL

Mr. Usman Khalil received a MS in Computer Sciences and a B.Sc. in Information Technology from University Brunei Darussalam (UBD), Brunei Darussalam, in 2019 and CECOS University, Pakistan in 2004 respectively. Currently, he is a Ph.D. scholar in Mathematical and Computing Sciences, Department of Digital Sciences at the University Brunei Darussalam. Previously, he has been working as a Telecom professional with more than fourteen years of professional experience in various telecom companies in Pakistan and Malaysia. His research interests focus on the exploratory pattern recognition algorithms for the analysis, data mining, integration of the Internet of Things with cloud computing, and the integration of the Internet of Things with blockchain.

DR. OWAIS AHMED MALIK

Dr. Owais Ahmed Malik is an Assistant Professor in Mathematical and Computing Sciences, Department of Digital Sciences, University Brunei Darussalam (UBD). He completed his Ph.D. in computer science from University Brunei Darussalam (2015), MS in Computer Science from KFUPM, Saudi Arabia (2002) and is a computer system engineering from NED, Pakistan (1998). Dr Owais has more than ten years of progressive experience in academia and research in the field of computer science and engineering. He has been teaching various undergraduate courses including machine learning, data mining, machine perception, programming fundamentals, design and analysis of algorithms, software engineering and operating system in different national/international universities. His research interest includes designing and exploring different intelligent/pattern recognition algorithms for the analysis and classification of biodiversity and cyber-security data, applied biomechanics, bio-signal processing and big data analytics. He has published a number of articles in internationally reputable journals and conferences. Email: owais.malik@ubd.edu.bn

DR. DAPHNE TECK CHING LAI



Dr. Daphne Lai is a Senior Assistant Professor at the Faculty of Science, she is also serving as a Director of the Institute of Applied Data Analytics at University Brunei Darussalam. She completed her PhD in computer science from University of Nottingham, UK, MS in Distributed systems and Networks, University of Kent, Canterbury, UK. and BSc in Computer Science, University of Strathclyde, UK. Her research interests lie in the areas of Data Mining, Artificial Intelligence and Metaheuristics. In recent years, she has been investigating on improving techniques for cluster analysis using evolutionary algorithms and machine learning. She is collaborating with researchers in several disciplines of health care, particularly cancer registry and cardiac rehabilitation, in geology and in traffic driving.