International Journal of Quantitative and Qualitative Research Methods
Vol.12, No.2, pp.44-57, 2024
Print ISSN 2056-3620 (Print)
Online ISSN 2056-3639(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Assessment of Agglomerative Clustering Techniques On Grouping of Malaria Infected Patients in Kaduna State, Nigeria

Omoha, Joy Omeyi[1]; Omole, Omotayo Omodele[2]; Abifade, Victor Oluwatobi[3]; Audi, Najib Isyaku[4]
Federal University of Health Sciences Otukpo; British Oak Global Institute; Bamidele Olumilua University of Education, Science and Technology Ikere Ekiti; Department of Statistics, Nasarawa State University, Keffi

doi: https://doi.org/10.37745/ijqqrm.13/vol12n24457 Published June 20, 2023

Citation: Omoha, J.O. Omole O.O., Abifade V.O., and Audi N.I. (2024) Assessment of Agglomerative Clustering Techniques On Grouping of Malaria Infected Patients in Kaduna State, Nigeria, *International Journal of Quantitative and Qualitative Research Methods*, Vol.12, No.2, pp.44-57

ABSTRACT: This study was carried out on assessment of agglomerative clustering techniques on grouping of malaria infected patients in kaduna state. The objective of this paper was to apply Agglomerative Hierarchical clustering techniques to group the variables, such that those similar to malaria infection will be identified. Reported cases of data relating to malaria cases were obtained on 1107 patients. Single linkage grouped sex and marital status at an early stage and joined by occupation at a farther distance which formed cluster I, while location and age joined at the same distance level to form cluster II. Complete Linkage grouped sex and marital status at a lower similarity which was later joined by occupation at a slightly higher distance to form cluster I. Average linkage, median. Centroid and ward's method also have sex grouped to marital status and joined closely by occupation to form cluster I. While age grouped with location to form cluster II. The location shows high similarity in all the method used which could be due to the swampy nature of the patient's environment. This paper concludes that the use of agglomerative clustering provides a suitable tool for assessing the diseases. It was recommended that the public health practitioners, policy makers, religious leaders and other stakeholders should use hierarchical clustering to develop strategies as a tool for disease control.

KEYWORD: Agglomerative Clustering, Malaria, Mortality rate.

INTRODUCTION

Hierarchical clustering methods generate a sequence of clusters solutions beginning with clusters containing a single object and combine objects until all objects from a single cluster: such method are called Agglomerative Hierarchical Clustering method. (Timm. 2002). Hierarchical clustering

is the sequence, conditionally partitioning of data from a single group containing all the observation to a partition where each group contains just one observation. The partitioning can be constructed in top down fashion, starting from the global cluster, which is divisive or commonly from the bottom up by merging groups together known as Agglomerative Clustering, (Nicholas et al, 2003).

Malaria is a mosquito-borne disease caused by a small one-celled parasite called plasmodium that infect and destroy red blood cells. Four different plasmodium can cause the disease in humans; plasmodium falciparum, plasmodium malariae, plasmodium ovale, and plasmodium vivax. Malaria is characterized by periodic bouts of severe chill and high fever, serious cases of malaria can resulting in death if left untreated. More than one million people die of the disease each year in Africa, (Microsoft Encarta, 2009). Malaria is the leading cause of morbidity and mortality in sub-Saharan Africa. As pregnant women and children are most at risk of severe malaria in areas of intense transmission of plasmodium faclapium.

During the past decades, have made significant progress in understanding the epidemiology of malaria worldwide and despite the improved understanding, the epidemic in Nigeria have continue to grow. People that have live malaria prone areas develop limited immunity to the disease especially infant and children which are most vulnerable.

LITERATURE REVIEW

The application of cluster analysis is inter disciplinary subject that spread to many numerous other fields like statistics, medicine, engineering, artificial intelligence, geology, biology, computer sciences, psychology and physiology, amongst others. The large number of application ranging from the classical ones such as automatic character recognition and medical diagnosis to the most recent one in data mining (such as credit cards scoring, consumer sales analysis, credit card analysis among others). Have attracted considerable research effort with many methods developed and advances made (Anderson, 1996).

Basically, multivariate analysis techniques are most commonly employed for developing taxonomies or systems of classification to investigate useful ways to conceptualize or group items, used to generate hypothesis and to test the hypothesis. Real world situation variable is measurement on a patient made in order to identify a disease (diagnosis). Measurement on patient in order to predict the likely outcome of (prognosis): measurement on weather variables (for forecasting prediction) among others (Srivastava and Khatvi, 1979).

Asoka *et al* (1991). Observed that malaria infection occurs most frequently among clustered population of people living in mud. Incomplete thatched roofs houses as against well constructed houses.Clustering is also the unsupervised classification of pattern (observation, data items. Or feature vector) into groups (clusters) as pattern within a valid clustering are more similar to each other than they are to a pattern belonging to different clusters, (Jain and Dubes. 1988a).

Clustering analysis have also been used in monkey model for malaria showing the similarities in gene expression as greater at similar stages for closely related once than those with far distance.(Joni *et al*. 2004). Cluster analysis is a technique which allows the identification of groups, or cluster of similar objects in multi-dimensional space. Ensemble clustering methods have also become increasingly important to ease the task of choosing the most appropriate cluster algorithms for a particular data analysis problem. (Michael et al, 2007). Although methods in these two groups have proved to be very effective and efficient, they generally depend on providing prior knowledge or information of the exact number of clusters for each dataset to be clustered and analyzed (Chang et al., 2010). More so, when dealing with real-world datasets, it is normal not to expect or have any prior information regarding the number of naturally occurring groups in the data objects (Liu et al., 2011). Clustering methods are used when a multi-dimensional space with a relatively high density of points is separated from other regions with high densities of points by regions with low densities of points, (Jain and Dubes, 1988b).

Cluster analysis attempts to identify the observation vector that are similar and group them into clusters, many techniques used an index of similarity or proximity between each pair of observation (Rancher, 1934). According to Johnson and Wichern (2001) most effort to produce a rather simple group structure from a complex data set required a measure of "closeness" or "similarity". When items (units or cases) are clustered. Proximity is usually indicated by some sort of distance. On the other hand, variables are usually grouped on the basis of correlation coefficient or like measures of association.

METHODOLOGY

In this paper, the method of analysis used in this research work is Agglomerative Hierarchical Clustering technique which emphasis on *Single, Complete, Average, linkage, Median, Centroid and ward's methods.* The choice of Agglomerative Hierarchical Clustering technique is informed by the fact that they are generally suitable for natural clusters. (Everitt, 1974).

To implement the Single Linkage Method, one combines objects in clusters using minimum dissimilarity between clusters, letter r represent any element in cluster R,r ε R, and s be any element in cluster S, s ε S from step 3 of the Agglomerative Clustering algorithms, distance between R and S are calculated using

$$D(R)(S) = \min\{d_{rs} \mid r \in R, s \in S\} \quad \text{(i)}$$

The result of a Hierarchical Clustering procedure can be displayed graphically using a tree diagram also known as dendogram, which shows all steps in the hierarchical procedures including the distance at which clusters merged, (Timm, 2002).

Complete Linkage method is also called the farthest neighbor method, the distance between two clusters R and S is detained as the maximum distance between a point in R and in S

D(R)(S) = max$\{d_{rs} \mid r \in R, s \in S\}$ (ii)

At each step, the distance is found for every pair of clusters, and the two clusters with the smallest distance are merged, (Timm 2002).

In the average linkage approach, the distance between two clusters *A* and *B* is defined as the averages of the *nAnB* distances between the *n4* points in *A* and *B* points in *B;*

D(A,B) = $\frac{1}{nAnB} \sum_{i-1}^{nA} \sum_{j=1}^{nB} d\ (yi, yj)$ (iii)

Where the sum is over all *yi* in *A* and all *yj* in *B*. at each step, we join two clusters with the smallest distance.

Ward's Method, also called the incremental sum of squares method uses within cluster (squared) distance (Ward 1963). If RS is cluster obtained by combining clusters R & S than the sums of within distance are.

$SSE_R = \sum_{l-1}^{nR} (y_i - y_R)\ (y_i - y_R)$ (iv)

$SSE_S = \sum_{i-1}^{nS} (y_i - y_S)(y_i - y_S)$

$SSE_{RS} = \sum_{i=1}^{nRS} (y_i - y_{RS})(y_i - y_{RS})$ (v)

Where $\bar{Y}_{RS} = (n_R\ y_R + n_S\ y)/(n_R + n_S), n_R n_S,$ and $n_{RS} = n_R + n_S$ are numbers of points in Rs, and Rs, respectively (Rencher, 1934).

The Median Method is called weighted version of centroid when a small cluster is joined to a larger one, the centroid of the result will closed to the centroid of the large cluster. It defines of clusters to be the distances between median of the clusters.

Centroid Methods calculate the proximity between two clusters by calculating the distance between the centroids of clusters. These techniques may seem similar to Kmeans.

Distance Measure Used for the Study

A common distance function used is the Euclidean distance between two vectors or p-dimensional observations (item) X= $(x_1 . x_{\ 2} \ldots \ldots . x_p)^1$ and Y= $(y_1 . y_{\ 2} \ldots \ldots . y_p)^1$, as defined as (Rencher 1934)

d(X – Y) = $\sqrt{(X-Y)^1\ (X-Y)}$

$$= \sqrt{\sum_{j=i}^{p}(x_j - y_j)^2} \quad \text{(vi)}$$

RESULTS AND DISCUSSION

Single Linkage Agglomeration Schedule & Proximity Matrix

Single Linkage Method clustering output displayed in Table 4.1 below describes the agglomeration schedule. The column under the "Cluster Combined" part of the "Agglomeration Schedule" Table shows which variables or clusters combined at each stage with respect to the malaria variables. This is reaffirmed by the proximity matrix in Table 4.2 and the icicle plot in Table 4.3 below. Here variables are merged based on the nearest and highest similarity levels. The nearest distance observed in Table 4.1 is 634.00 which merge the pair of variables, sex and marital status of the malaria cases. The next highest similarity level is occupation joined to marital status and sex with a distance of 121,983.00, while age and location stood on their own. When clusters are joined, the "Coefficient" value and the proximity distance depend on the linkage method used. At the final stage of cluster formation, occupation joined the first cluster of sex and marital status to form a group. It is often easier to follow how groups and individual cases join together in this process in a dendogram, which is displayed in below.

Single Linkage

Table 4.1: Agglomeration Schedule

Stage	Cluster Combined		Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Coefficients	Cluster 2	Cluster 1	
1	1	3	634.000	0	0	
2	2					
3	1	4	121983.000	1	0	4
4	1	5	273205.000	0	0	4
	1	2	276835.000	2	3	0

Table 4.2: Single Linkage Proximity Matrix

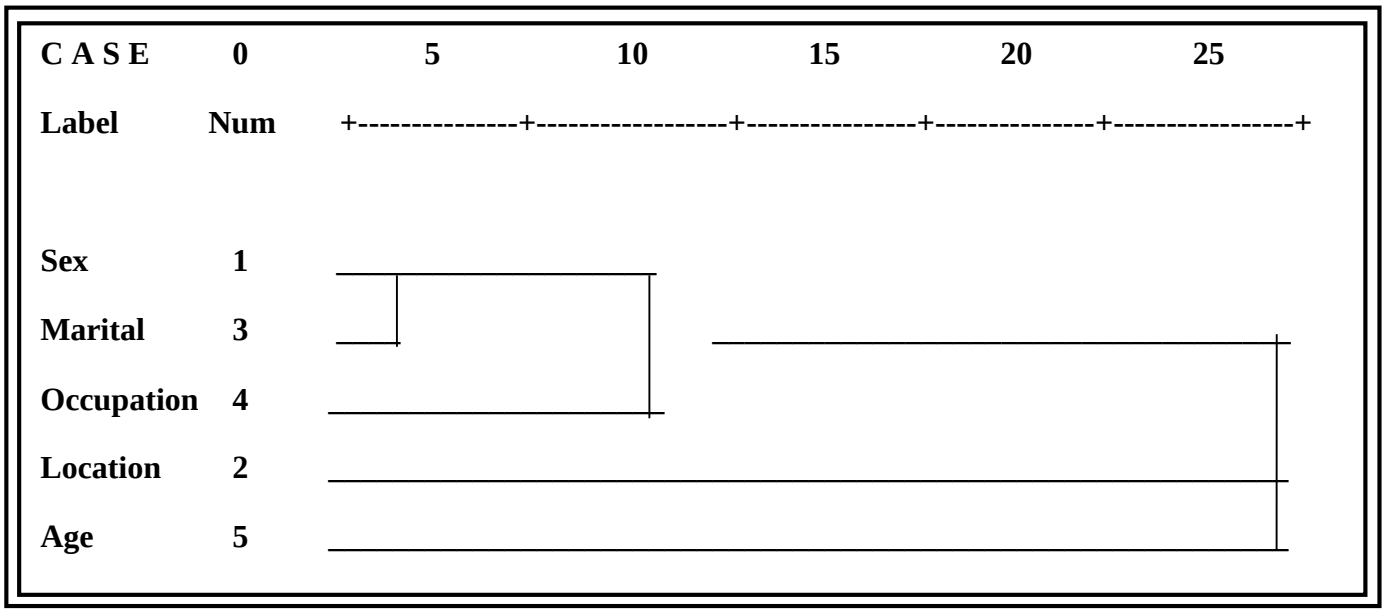
Case	Matrix File Input				
	Sex	Location	Marital Status	Occupation	Age
Sex	0.000	522038.000	634.000	124747.000	681023.000
Location	522038.000	0.000	514554.000	276835.000	273205.000
Marital Status	634.000	514554.000	0.000	121983.000	675543.000
Occupation	681023.000	276835.000	121983.000	0.000	333816.000
Age	681023.000	273205.000	675543.000	332816.000	0.000

Table 4.3: Single Linkage Horizontal Icicle

Case	Number Of clusters			
	1	2	3	4
Age	X	X	X	X
	X	X		
Location	X	X	X	X
	X			
Occupation	X	X	X	X
	X	X	X	
Marital Status	X	X	X	X
	X	X	X	X
Sex	X	X	X	X

International Journal of Quantitative and Qualitative Research Methods
Vol.12, No.2, pp.44-57, 2024
Print ISSN 2056-3620 (Print)
Online ISSN 2056-3639(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Dendrogram I



Complete Linkage Method clustering output displayed in Table 4.4 above describes the agglomeration schedule. This method proceeds in the same manner as the single linkage only that complete linkage ensures the variables in clusters are within some maximum distance (or minimum similarity) to each other. This is reaffirmed by the proximity matrix

in Table 4.3 and the icicle plot above. The scale distance increases from 634.000 up to 681023.000. The distribution of variable shows that the two closest variables to form a cluster in marital status and sex at a distance of 634.00. The next cluster is age and location, as occupation stood on its own until the final stage where it was merged with marital status and sex. In agglomeration clustering, when clusters are joined, the "Coefficient" value and the proximity distance depend on the linkage method used. It is often easier to follow how groups and individual cases join together in this process in a dendrogram.

Complete Linkage

Table 4.4 Agglomeration Schedule

Stage	Cluster Combined			Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Coefficients	Cluster 2	Cluster 1	
1	1	3	634.000	0	0	2
2	1	4	124747.000	1	0	4
3	2	5	273205.000	0	0	4
4	1	2	681023.000	2	3	0

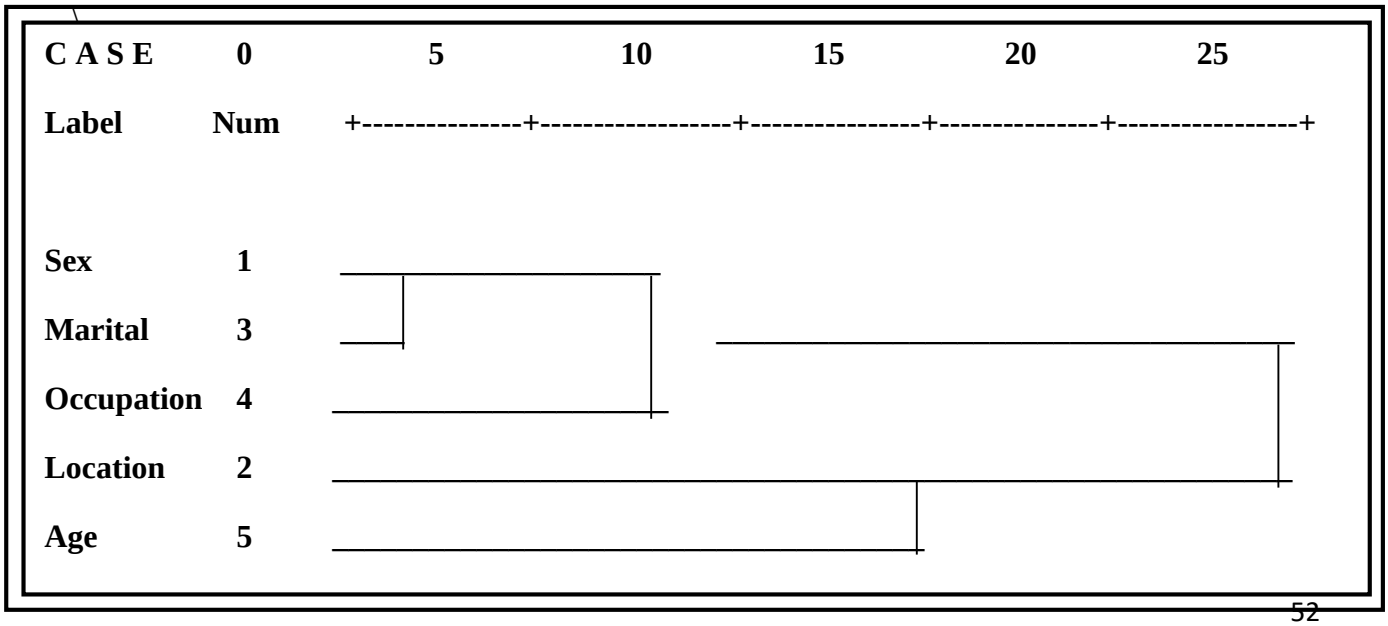
Table 4.5: Complete Linkage Proximity Matrix

Case	Matrix File Input				
	Sex	Location	Marital Status	Occupation	Age
Sex	0.000	522038.000	634.000	124747.000	681023.000
Location	522038.000	0.000	514554.000	276835.000	273205.000
Marital Status	634.000	514554.000	0.000	121983.000	675543.000
Occupation	124747.000	276835.000	121983.000	0.000	332816.000
Age	681023.000	273205.000	675543.000	332816.000	0.000

Table 4.6: Complete Horizontal Icicle

Case	Number of Clusters			
	1	2	3	4
Age	X X	X X	X	X
Location	X X	X	X	X
Occupation	X X	X X	X X	X
Marital Status	X X	X X	X X	X X
Sex	X	X	X	X

Dendrogram II



Average Linkage Method merge subgroups based on the smallest distance but define distance as average of all pairs of items, one from each group. The clustering output displayed in Table 4.7 above describes the agglomeration schedule. This is reaffirmed by the proximity matrix in Table 4.8 and the icicle plot in Table 4.9 respectively above. Firstly, the Euclidean distance between marital status and sex is 634.00 which are smaller than the distance between any other pair of malaria variables. Secondly, occupation is joined to marital status and sex with a distance of 123,365.00 and so on. The final formations have occupation joined to marital status and sex, while age is joined to location. In agglomeration clustering, when clusters are joined, the "Coefficient" value and the proximity distance depend on the linkage method used. Here with Average Linkage Method, the distance between marital Status and sex is 634.00 units. It is often easier to follow how groups and individual cases join together in this process in a *dendrogram*, which is displayed in figure 3 below.

Average Linkage

Table 4.7 Agglomeration Schedule

Stage	Cluster Combined		Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Coefficients	Cluster 2	Cluster 1	
1	1	3	634.000	0	0	2
2	1	4	123365.000	1	0	4
3	1	2	273205.000	0	0	4
4	1	5	500468.167	2	3	0

Table 4.8: Average Linkage Proximity Matrix

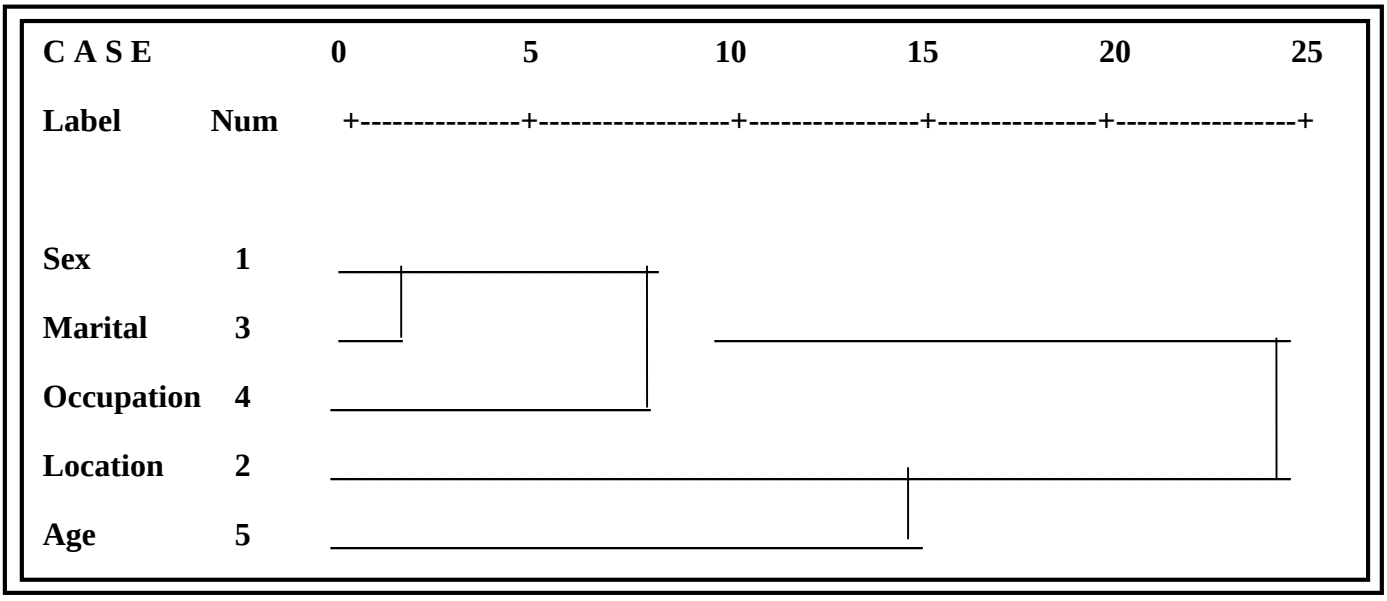
Case	Matrix File Input				
	Sex	Location	Marital Status	Occupation	Age
Sex	0.000	522038.000	634.000	124747.000	681023.000
Location	522038.000	0.000	514554.000	276835.000	273205.000
Marital Status	634.000	514554.000	0.000	121983.000	675543.000
Occupation	124747.000	276835.000	121983.000	0.000	332816.000
Age	681023.000	273205.000	675543.000	332816.000	0.000

International Journal of Quantitative and Qualitative Research Methods
Vol.12, No.2, pp.44-57, 2024
Print ISSN 2056-3620 (Print)
Online ISSN 2056-3639(Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Table 4.9: Average Linkage Horizaontal Icicle

Case	Number of clusters			
	1	2	3	4
Age	X	X	X	X
	X	X		
Location	X	X	X	X
	X			
Occupation	X	X	X	X
	X	X	X	
Marital Status	X	X	X	X
	X	X	X	X
Sex	X	X	X	X

Dendrogram III



Unlike single and compete linkage, the centroid fuse group according to the distance between their centroids. The clustering output displayed in table 4.10 above described the agglomeration schedule. This is reaffirmed by the matrix

in Table 4.8 and the icicle plot in Table 4.9 respectively above. Marital status and sex are clustered at 634.00 which are smaller than the distance between any other pair of malaria variables. Second, occupation is joined to marital status and sex with a distance of 123,206.00 and so on. From the results, it can be deduced that the centroid linkage method produces similar results with the average linkage method. In agglomeration clustering, when clusters are joined, the "Coefficient" value and the proximity distance depend on the linkage method used. Here the Centroid Linkage Method, the distance between marital status and sex is 634.00 units. The *dendrogram* is displayed in Figure 4 below.

Cluster Formation by Various Method

The numbers of clusters formed by the various methods are same with varying distance/similarity level. Comparing the various methods from one to another indicates that they all formed two main clusters. The dendrogram produces by the six methods viz: single linkage, compete linkage, average linkage, median method, ward's method and centroid method shows small changes in levels. All the methods grouped sex and marital status in the first cluster with occupation joining at a later distance, it is pertinent to mention, that the clustering of malaria variables using all the agglomeration methods implies that three malaria variables – sex, marital status and occupation could be clustered together while age and location could also form another cluster.

Sex, marital status and occupation formed are common to all in cluster I, while cluster II is dominated by location and age respectively. It was also observed that similarity exist between age and location. We have been able to identify the variables that have similar and dissimilar pattern. Based on the analysis carried out the most similar and peculiar variable to all methods are sex, marital status, age and location. The study has shown that the six agglomerative methods used were efficient, because the variables grouped are almost the same and each method grouped variables at different similarity or distance level. However, they form same numbers of clusters. It also shows that multivariate method is another tool for assessing the level of infection of the disease. The result provides the information on the various agglomerative clustering methods based on the clusters formed by each method. The result provide the inter-correlation between the various methods.

CONCLUSION

In the analysis carried out, we have been able to show the natural grouping of malaria variables considered in this research work using single linkage, complete linkage, average linkage, median.

Recommendations

Based on the result obtained from the analysis, it is recommended that public enlightens programme should be step down to the vulnerable area especially grassroots to enhance healthy living. Attention should be focused on pregnant women and children especially in the use of insecticides and mosquito nets. Also malaria vaccine should be made available for children as this will reduce mortality rate. The federal government should develop and implement some health policy in addition to the existing one to help prevent the disease epidemics. The government should also make fund available to enhance researchers make new discoveries that will eradicate the disease nationwide. The public health practitioners, policy makers, religious leaders and other stakeholders should use hierarchical clustering to develop strategies as a tool for disease control. Finally, to adequately address malaria spread, it is recommended that a population-based survey of malaria with emphasis on location, standard of living, health history, numbers of mosquito net in a home, and so on.

References

Anderson, T.W. (1996) "Multivariate Analysis", *Journal Of Statistical Science:* 11(1). 20-34.

Asoka, C.G., Carter, R.. Mendis, C. And Mendis, K. N. (1991) "Clustering Of Malaria Infections Within An Endemic Population. *Journal Of Medical Sciences,* 3, 77-85.

Dubien, J.L. And Warden, W.D. (1979), "A Mathematical Comparison Of Members Of An Infinite Family Of Agglomerative Clustering Algorithms," *Canadian Journal Of Statistics, 7,29-38.*

Everitt, B.S., Landau, S., Leeses, M. And Stahl, D. (2011), *Cluster Analysis* 5th Ed, John Wiley & Sons, Ltd, Chichester, UK.

El-Hamdouchi, A. And Willet, .P. (1989). Comparison Of Hierarchical Agglomerative Clustering Methods For Document Retrieval. *The Computer Journal, 32(3), 220-227.*

Florence, C. (1988). "Multiple Sequence Alignment With Hierarchical Clustering", *Journal Of American Statistical Association,* 16.(22), 81-90.

Hartigan, J.A (1972). Direct Clustering Of Data Matrix, *Journal Of American 'Statistical Association; 28, 123-129.*

International Association For Medical Assistance To Travelers (IAMAT) 2012 March. *"How To Protect Yourself Against Malaria",* Website: Www.Iamat.Org.

Jason A.Y And Winzeler, E.A. (2005), "Using Expression Information To Discover New Drug And Vaccine Targets In The Malaria Parasite Plasmodium Falcapium", *Journal Of Infectious Diseases.* 6(1) 17-26.

Johnson, R.A And Wichern D.W (2001), "Applied Multivariate Statistics Analysis *6th Ed,* Pearson Prentice Hall, 223-256.

Joni .Y. Amber C.R, Tereance A.M (2004) Transcriptome Profiles Of Host Gene Expression In Monkey Model Of Human Malaria. *Journal Of Infectious Diseases,* 191,(3) 400-409.

Lance, G.N And Williams, W.T (1967), The Africa Theory Of Classificatory Sorting Strategies: Hierarchical Systems, *Computer Science Journal,* 9, 60-64.

Linda .W. (2008) *Sampling Methods* Excerpt From The *Certified Software Quality Engineer Handbook, 4,53-6, New York.*

Michael .H, Stephen .S, And Xiohui, L (2007) "Optimal Search Space For Clustering Gene Expression Data Via Consensus" *Journal Of Computational Biology*, 14,1327-1341.

Microsoft Encarta And Student Program Manager, (2009), *"History Of Malaria".* One Microsoft Way Redmond, WA 98052-6399 USA.

Ministry Of Health Tanzania National Guideline (2001) For Malaria Diagnosis And Treatment, 1st Edition, United Republic Of Tanzania: Ministry Of Health.

Nicholas, C. Marc, T And Jacob .K. (2003). "Data Driven Similarity Measures For K-Means Like Clustering Algorithms. *Computer Science Journal*, 2, 12-16.

Rosie, C. (2007) "Statistics Cluster Analysis" Mathematics Learning Support Centre, USA. (2), 35-50. Www.Presselosjen.No/Content/Text.

Srivastava, M.S. And Khatri, C.G. (1979). "An Introduction To Multivariate Statistics" 2nd Ed. New York: North Holland.

Timm. N.H (1975) "Multivariate *Analysis With Application In Education And Psychology"* Wadsworth Publishing Company; California.

Treatment Of Malaria (Guidelines For Clinicians) Www.Cdc.Gov/Malaria/Clinicans.Httm# Report June 28th 2004, 1-4.

Ward, J.H. (1963), "Hierarchical Grouping To Optimize An Objective Function, *"Journal Of American Statistical Association:* (58), 236-244.

Wolfgang .H And Leopold, S. (2003) "Applied Multivariate Statistical" Butterworth-Heinemann Ltd, United Kingdom.