Biomedical Statistics and Informatics
2017; 2(3): 122-127
http://www.sciencepublishinggroup.com/j/bsi
doi: 10.11648/j.bsi.20170203.16

Classification of Some Seasonal Diseases: A Hierarchical Clustering Approach

Samson Agboola, Mataimaki Benard Joel

Department of Statistics, Faculty of Physical Science, Ahmadu Bello University, Zaria, Nigeria

Email address:
abuagboola@gmail.com (S. Agboola), manjoello@yahoo.com (M. B. Joel)

To cite this article:
Samson Agboola, Mataimaki Benard Joel. Classification of Some Seasonal Diseases: A Hierarchical Clustering Approach. *Biomedical Statistics and Informatics*. Vol. 2, No. 3, 2017, pp. 122-127. doi: 10.11648/j.bsi.20170203.16

Received: July 22, 2017; **Accepted**: August 2, 2017; **Published**: September 4, 2017

Abstract: This study compared six (6) agglomerative hierarchical clustering techniques namely Single-linkage, Complete-linkage, Centroid hierarchical, group average linkage, median hierarchical and ward's minimum variance on some seasonal diseases to know which technique is most appropriate for classification. These seasonal diseases where gotten from five (5) different hospitals namely; Jamaa, Salama, Almadina, Gambo Sawaba and St Lukes Hospitals in Zaria. The Root Mean Square Distance Between Observation (RMS-DBO) which gives the best technique (s) for classification showed that the single-linkage and complete-linkage was the best techniques for the classification of the diseases. The results were calculated using R and SAS packages. The study achieves the best clustering technique for the classification of the studied seasonal diseases.

Keywords: Hierarchical, Clustering, Diseases, Classification, RMS-DBO, Techniques

1. Introduction

Many common diseases, such as the flu and cardiovascular disease, increase markedly in dry (winter) and dip in wet (summer). The seasonal patterns and climatic sensitivities of many transmittable diseases are well known; the important contemporary concern is the extent to which changes in disease patterns will occur under the conditions of global climate change. Seasonal cycles of infectious diseases have been variously attributed to changes in atmospheric conditions, the occurrence of the pathogen, or the behavior of the host. Detection of outbreaks is an important part of disease surveillance.

Anderson (1996) stated that, the field of multivariate analysis consists of those statistical techniques that consider two or more related random variables as a single entity and attempt to produce an overall result taking the relationship among the variables into account. (Morrison 1990) stated that multivariate statistical analysis is concerned with data collected on several dimensions of the same individual; such observations are common in the social, behavioral life and medical science. It therefore helps the researcher to summarize the data and reduce the number of variables necessary to describe it. Morison (1990) further stated that, as a result of the complexity of the variables, multivariate techniques such as cluster analysis will be used for classification.

Cluster analysis was defined by (Hartigan 1972) as one tool of exploratory data analysis that attempts to assess the interaction among patterns within clusters are more similar to each other, than those whose patterns belong to different clusters. Cluster analysis is multivariate analysis technique seeks to organize information about variables so that relatively homogenous "groups" or "clusters" can be formed. The clusters formed are internally homogenous and externally heterogeneous. That is, variability within a group is minimum and variability between groups is maximum. The results of cluster analysis can be used to initiate hypothesis about the data, to classify new data, to test for homogeneity of the data and to compress the data (Ryan 1977). (Cornell, *et al*. 2007) used cluster analysis in clustering binary data from multimorbidity clusters that is clustering binary data from a Large administrative medical database. His main aim was to describe and illustrate the application of cluster analysis to identify clinical relevant multimorbidity groups. (Dauda, *et al*. 2011) worked on Monetering of infectious Diseases in Katsina and Daura Zones of Katsina State, A clustering Analysis. In their work, they used the single linkage, Centroid method, Complete Linkage and ward method to classify the infectious diseases into groups such that the diseases with

similar degree of prevalence were identified. Their finding was that the use of clustering methods provides a suitable tool for assessing the level of infections of the diseases. Gulumbe *et al*. (2008) applied hierarchical clustering techniques to partition the set of variables into groups, such that those are similar with respect to HIV/AIDS. Infections were identified and two main clusters were observed. The implication of the cluster formation shows that HIV/AIDS infection is more prevalent among married women as in single and ward's linkage methods. It also shows that the disease affect mostly the working class aged from 15 to 39 as grouped by complete linkage method. The relationship between the various methods used and clusters formed with respect to the variable grouped were found to be consistent using chi square test for independence. (Nwabueze, 2013) used the agglomerative hierarchical clustering to group cassava mosaic disease-resistant varieties cultivated by National Root crops research institute, Umudike, Nigeria. Her findings showed that the agglomerative hierarchical schedule showed that the varieties could be placed in five distinct groups. The dendrogram of the study which showed the relative size of the proximity coefficient at which cases were combined was obtained. (Jain and Chaturvedi 2014), used Hierarchical Clustering technique in clustering of result to obtain an approach for cancer Disease detection. This approached used by Jain and Chaturvedi help a lot to understand the result provided by the proposed system of quantum computing based technique for diagnoses of cancer. The main reason for having many clustering methods is the fact that the notion of 'cluster' is not precisely defined (Estivill-Castro, 2000). Consequently many clustering methods have been developed, each of which uses a different induction principle. Fraley and Raftery (1998) suggest dividing the clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber (2001) suggest categorizing the methods into additional three main categories: density-based methods, model-based clustering and grid-based methods. Although these types are not necessarily mutually exclusive and several clustering techniques could be placed in more than one category. For this research work, the Hierarchical clustering techniques will be used and the methods described. The choice of hierarchical clustering technique is informed by the fact that they are generally suitable for natural cluster.

Hierarchical clustering combines cases into homogeneous clusters by merging them together one at a time in a series of sequential steps (Blei & Lafferty, 2009). Nonhierarchical techniques (e.g., k-means clustering) first establish an initial set of cluster means and then assign each case to the closest cluster mean (Morissette & Chartier, 2013). Hierarchical cluster analysis can be conceptualized as being agglomerative or divisive. Agglomerative hierarchical clustering separates each case into its own individual cluster in the first step so that the initial number of clusters equals the total number of cases (Norusis, 2010). Divisive hierarchical clustering works in the reverse manner with every case starting in one large cluster and gradually being separated into groups of clusters until each case is in an individual cluster. This latter technique, divisive clustering, is rarely utilized because of its heavy computational load. (Wilmink & Uytterschaut, 1984). Hands and Everitt (1987) examined five hierarchical clustering techniques on multivariate binary data, comparing their abilities to recover the original clustering structure. They controlled various factors including the number of groups, number of variables, proportion of observations in each group, and group-membership probabilities. All of the clustering techniques in the study were hierarchical: single linkage, complete linkage, group average, centroid, and Ward's method. Johnson and Wichern (2002) informally contrast various hierarchical clustering methods, without specifying a particular data type (the implication is that they are considering multivariate continuous data). For instance, they note single linkage tends to perform poorly with clusters that are truly elliptical. This problem occurs especially when clusters are elliptical when plotted, but the ends of the clusters are geometrically close. This causes certain points to be clustered together early in a hierarchical algorithm when in fact they should be placed with the rest of their own ellipse. Tarpey (2007) compared several clustering methods for functional data. He focused on k-means clustering and examined the effect on the clustering outcomes based on how the observed data were smoothed (using the raw data, and data smoothed via a B-spline basis, Fourier basis, and power basis, respectively). He concluded that the results of clustering functional data depend on how well the smooth curves fit the raw data, but that the choice of best smoothing method depends on the true mean curve of each cluster.

An imperative deduction arising from most of these studies was that the advantage of one technique over others was not uniform, but rather depended greatly on the form of the data.

2. Method

2.1. Data

The data were obtained from Hajiya Gambo Sawaba General Hospital, Salama Specialist Hospital, Jama'a Hospital, Almadina Specialist Hospital and St. Lukes hospital Zaria, Kaduna State, Nigeria. The data also indicates gender, age, occupation, diagnosis, date and the year the patients reported and the addresses of the Patients. The seasonal diseases studied were pneumonia, Typhoid, measles, meningitis and Diarrheas. The study covered a period of ten (10) years from 2005 to 2014 in which over 2,162 patients visited the five hospitals. 1030 of the patients were females and 1132 where males. All the patients were admitted with either diarrhea, measles, meningitis, pneumonia and typhoid fever.

2.2. Study Area

The study area is Zaria, a major city in Kaduna State in

Northern Nigeria as well as being a Local Government Area. Formerly known as Zazzau, it was one of the original seven Hausa city-states. However, human settlement predates the rise of Zaria as the region, like some of its neighbors, had a history of sedentary Hausa settlement, with institutional but pre-capitalist market exchange and farming. The 2006 Census population showed that Zaria's population was 408,198. Zaria has a total Area of about $300km^2$ (100 sq mi). The old part of the city, known as Birnin Zaria or Zaria-City, was originally surrounded by walls, which now have been mostly removed. The Emir's palace is located in the old city. In the old city and the adjacent Tudun Wada neighbourhood people typically reside in traditional compounds. These two neighborhoods are predominately occupied by the indigenous Hausa. The neighborhoods of Samaru and Sabon Gari are predominately occupied by Nigerians of southern origin, such as the Igbo. The largest marketplace is in Sabon Gari. Other more recent neighborhoods include: Danmagaji/Wusasa, PZ, Kongo, GRA-Zaria, Hanwa, Bassawa, Lowcost Kofan-Gayan and Shikka. Its coordinates is $11^0 04' N\ 7^0 42' E$. 1030 of the patients were females and 1132 where males. All the patients were admitted with either diarrhea, measles, meningitis, pneumonia and typhoid fever. From our records, it was established that diarrhea patients heard the highest admission into the various hospitals with a number of 743 followed closely with typhoid fever patients with 733, measles (255), meningitis (207) and pneumonia patients were 224. All of which are duly due to the harsh conditions of our weather conditions in Nigeria.

2.3. Materials and Methodology

The method of analysis used is the agglomerative hierarchical clustering technique where emphasis is on single-link, complete-link, centroid, average, median and ward's linkage method. The choice of agglomerative hierarchical clustering technique is informed by the fact that they are generally suitable for natural cluster; no need to specify the number of clusters in advance, its structures maps nicely onto human intuition for some domains and interpretation of the result is (very) subjective. The techniques performs reasonably well when clusters are clearly separated (Everitt, 1974). In general, the six linkage methods will be used as this will help to prevent misleading solutions being accepted. More so, measurement made for all the linkage methods are based Root Mean Square Distance Between Observation (RMS-DBO). However differences in the linkage methods are due to difference in defining distance (similarity) between groups for each method (Everitt, 1974).

2.4. Single Linkage Method

Single linkage method also referred to as nearest neighbour or minimum method that considers the distance between two clusters to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, the similarity between a pair of clusters is considered to be equal to the greatest similarity from any member of one cluster to any member of the other cluster (Sneath and Sokal, 1973).

If A and B, is the distance between their closest variables,

$$d_{AB} = \underset{i\varepsilon A, j\varepsilon B}{\boldsymbol{min}}\, d_{ij} \tag{1}$$

In a matrix, the element in i^{th} row and j^{th} column gives the distance, d_{ij}, between individuals' i^{th} and j^{th} variables.

The result of single linkage clustering can be graphically displayed in the form of dendrogram, or tree diagram. The branches in the dendrogram represent clusters. The branches come together (merge) at nodes whose position along a distance (or similarity) axis indicate the level at which the fusion occur.

The distance between group i (containing n_i) observation and grouping j (containing n_j) is the least of $n_i n_j$ distance between element i and j

$$D_{k\,ij} = \frac{1}{2}\left[D_{ki} + D_{kj} - |D_{ki} - D_{kj}|\right] \tag{2}$$

2.5. Complete Linkage Method

Complete linkage method or furthest neighbor linkage method proceed in much the same manner with single linkage method with only one important exception. At each stage, distance between the two furthest points is considered. That is, at each stage, distance between clusters is determined by the distance between the two variables, one from each cluster that are further away distant. Thus, complete linkage method ensures that all variables in a cluster are within some maximum distance to each other.

If AB, is the distance between their furthest variables, then,

$$d_{AB} = \underset{i\varepsilon A, j\varepsilon B}{\boldsymbol{max}}\, d_{ij} \tag{3}$$

Where d_{AB} is the maximum distance between the paired variables AB with another cluster

When D_{ij} is the greatest of the distance between element of i and element j then

$$D_{k\,ij} = \frac{1}{2}\left[D_{ki} + D_{kj} + |D_{ki} - D_{kj}|\right] \tag{4}$$

2.6. Centroid Linkage Method

The centroid method of hierarchical cluster analysis is based on the distance between their centroid clusters. That is, the distance between groups is defined as the distance between the group centroids. The procedure is then to fuse groups according to distance between their centroids, the groups with the smallest distance being fused first. The centroid method can be used with both similarity and distance measures.

Let D_{ij} be the squared Euclidean distance between the centroids of group i and group j, then

$$D_{k\,ij} = \frac{n_i}{n_1 + n_j} D_{ki} + \frac{n_i}{n_i + n_j} D_{kj} - \frac{n_i n_j}{n_i + n_j} D_{ij} \tag{5}$$

2.7. Group Average Clustering

In the group average method, the distance between two clusters is defined to be the average of the dissimilarities between all pairs of individuals, one from each group.

$$D_{ij}\ AB = \frac{1}{n_i + n_j} = \sum_{i\varepsilon A, j\varepsilon B}\ d_{ij} \quad (6)$$

Let D_{ij} be the average of $n_i n_j$ inter-element distance, then

$$D_{k\ ij} = \frac{n_i}{n_i + n_j}\ D_{ki} + \frac{n_j}{n_i + n_j}\ D_{kj} \quad (7)$$

2.8. Median Clustering

This is also called un-weighted version of centroid clustering. When a small cluster is joined to a larger one, the centroid of the result will be close to the centroid of the large cluster. For some problems this may be a disadvantage. This measure attempts to overcome this by defining the distance between two clusters to be distance between the medians of the clusters.

Let D_{ij} be the distance between "median" of i and j, then

$$D_{k\ ij} = \frac{1}{2}\ D_{ki} + \frac{1}{2}\ D_{kj} - \frac{1}{4}\ D_{ij} \quad (8)$$

2.9. Ward's Linkage Method

This method of hierarchical cluster analysis is based on sum of squares between clusters summed over all variables. That is, ward's method attempt to minimize the sum of squares (SS) of any two (hypothetical) clusters that can be formed at each stage. The distance is the increase in total sum of squares which would result if the two clusters were combined and this is deviation of observation points from their cluster means.

The distance is computed as below;

$$\mathrm{d}(x_i, x_j) = \sum\sum\left(x_i - x_j\right)(x_i - x_j)^1 \quad (9)$$

so that, a distance matrix is form by $D=d_{ij}$ where x_i and x_j are the observation point of the variable x. at each step in the analysis, union of each possible pair of cluster is considered and two cluster whose fusion result in the minimum increase in the error sum of square are combined. The aim of ward's method is to minimize the total within-group sum of squares.

3. Result

The analysis is aimed at showing which six (6) Hierarchical clustering technique (s) is most suitable for classification. The cluster procedure showed the eigenvalues of all the clustering techniques, with the difference and proportion. The Root Mean Square- Total Sample Standard Deviation (RMS-TSSD) was derived as well as the Root Mean Square Distance Between Observation (RMS-DBO).

The CLUSTER Procedure

3.1. Average Linkage Cluster Analysis

Eigenvalues of the Covariance Matrix
Eigenvalue Difference Proportion Cumulative
1 382.736724 374.484268 0.9783 0.9783
2 8.252457 8.023283 0.0211 0.9994
3 0.229173 0.0006 1.0000
Root-Mean-Square Total-Sample Standard Deviation = 11.41955
Root-Mean-Square Distance Between Observations = 27.97207

3.2. Centroid Hierarchical Cluster Analysis

Eigenvalues of the Covariance Matrix
Eigenvalue Difference Proportion Cumulative
1 382.736724 374.484268 0.9783 0.9783
2 8.252457 8.023283 0.0211 0.9994
3 0.229173 0.0006 1.0000
Root-Mean-Square Total-Sample Standard Deviation = 11.41955
Root-Mean-Square Distance Between Observations = 27.97207

3.3. Single Linkage Cluster Analysis

Eigenvalues of the Covariance Matrix
Eigenvalue Difference Proportion Cumulative
1 382.736724 374.484268 0.9783 0.9783
2 8.252457 8.023283 0.0211 0.9994
3 0.229173 0.0006 1.0000
Root-Mean-Square Total-Sample Standard Deviation = 11.41955
Root-Mean Distance Between Observations = 22.14743

3.4. Ward's Minimum Variance Cluster Analysis

Eigenvalues of the Covariance Matrix
Eigenvalue Difference Proportion Cumulative
1 382.736724 374.484268 0.9783 0.9783
2 8.252457 8.023283 0.0211 0.9994
3 0.229173 0.0006 1.0000
Root-Mean-Square Total-Sample Standard Deviation = 11.41955
Root-Mean-Square Distance Between Observations = 27.97207

3.5. Median Hierarchical Cluster Analysis

Eigenvalues of the Covariance Matrix
Eigenvalue Difference Proportion Cumulative
1 382.736724 374.484268 0.9783 0.9783
2 8.252457 8.023283 0.0211 0.9994
3 0.229173 0.0006 1.0000
Root-Mean-Square Total-Sample Standard Deviation = 11.41955
Root-Mean-Square Distance Between Observations = 27.97207

3.6. Complete Linkage Cluster Analysis

Eigenvalues of the Covariance Matrix
Eigenvalue Difference Proportion Cumulative
1 382.736724 374.484268 0.9783 0.9783
2 8.252457 8.023283 0.0211 0.9994
3 0.229173 0.0006 1.0000
Root-Mean-Square Total-Sample Standard Deviation = 11.41955
Root-Mean Distance Between Observations = 22.14743

4. Discussion

It was discovered that the eigenvalue (382.736724, 8.252457, 0.229173) of all the clustering techniques were the same. The difference (374.484268, 8.023283) and the proportion (0.9783, 0.0211, and 0.0006) were also the same. The Root Mean Square Total sample standard deviation (11.41955) of the clustering techniques was also the same. The Root mean square distance between observation (RMS-DBO) varied. This is because RMS-DBO gives the best technique (s) suitable for classification. Our analysis showed that the single-Linkage and the Complete Linkage cluster analysis are most suitable for classification because the result showed 22.14743 which is less than the other clustering techniques with 27.14743.

5. Conclusion

This research work was aimed at using Hierarchical clustering technique (HCT) for spatial classification of some seasonal diseases. We applied six (6) agglomerative clustering techniques to five (5) seasonal diseases. The Root Mean Square Distance Between Observation which shows the best model for classification showed that the single-linkage and complete-linkage was the best technique for the classification of the diseases with 22.14743 whereas the group average, centroid, median and Ward methods was 27.97207. From our records, it was established that diarrhea patients heard the highest admission into the various hospitals with a number of 743 followed closely with typhoid fever patients with 733, measles (255), meningitis (207) and pneumonia patients were 224. All of which are duly due to the harsh conditions of our weather conditions in Nigeria.

Based on the result obtained from the analysis of demographic and spatial data, diarrhea being the most prevalent disease followed by typhoid in the study area shows that the government put in place strategies of reducing the menace or occurrences of these diseases. More enlightenment campaign should be made by non-governmental organisations (NGOs) such as World Health Organization (WHO) to help fight the occurrences of these diseases.

References

[1] Anderson, T. W. "Fisher and Multivariate Analysis." (Statistical Science journal) 11, no. 1, 20-34 (1996).

[2] Blei, D. & Lafferty, J. (2009). Topic models. In A. Srivastava and M. Sahami (Eds.), Text Mining: Classification, Clustering, and Applications (pp. 71-94). Boca Raton, FL: Taylor & Francis Group.

[3] Cornell, E. John, et al. "Multimorbidity Clusters: Clustering binary data from multimorbidity cluster: Clustering binary data from a large administrative database." *Applied Multivariate Research*, 2007: 163-182.

[4] Dauda, U, S. U Gulumbe, M Yakubu, and L. K Ibrahim. "Monetering of Infectious Diseases in Katsina and Daura Zones of Katsina State: A Clustering Analysis." *Nigerian Journal of Basic and Applied Science*, 2011: 31-42.

[5] Everitt, B. S. "Cluster Analysis", Heinemann Educational Book Ltd, UK. 1974.

[6] Fraley C. and Raftery A. E., "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis", Technical Report No. 329. Department of Statistics University of Washington, 1998.

[7] Hands, S. and Everitt, B. (1987). A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. Multivariate Behavioral Research, 22, 235–243.

[8] Norusis, M. J. (2010). Chapter 16: Cluster analysis. PASW Statistics 18 Statistical Procedures Companion (pp. 361-391). Upper Saddle River, NJ: Prentice Hall.

[9] Nwabueze, Joy Chioma. "Statistical grouping of cassava mosaic disease-resistant varieties cultivated by the National Root Crops Research Institute, Umudike, Nigeria." *African Journal of Mathematics and Computer Science Research*, 2013: 26-34.

[10] Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

[11] Hartigan, J. A. "Direct Clustering of Data Matrix." (Journal of American statistical Association) 67, no. 123-129 (1972).

[12] Gulumbe, S. U., Bakar, A. B. and Dikko, H. G. (2008). Classification of some HIV/AIDS Variables, a multivariate approach. *Res. J. Sci.* 15: 24 – 30.

[13] Jain, Milan, and Setu Kumar Chaturvedi. "Quantum Computing Based Technique for Cancer Disease Detection System." *Journal of Computer Science & Systems Biology*, 2014: 9.

[14] Johnson, R. A, and D. W Wichern. "Applied Multivariate Statistical Analysis." *Prince Hall*, 2002.

[15] Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematical. Tutorials in Quantitative Methods for Psychology, 9 (1), 15-24.

[16] Morrison, D F. "Multivariate Statistical Method." (MC Craw Hill) 1990.

[17] Nwabueze, Joy Chioma. "Statistical grouping of cassava mosaic disease-resistant varieties cultivated by the National Root Crops Research Institute, Umudike, Nigeria." *African Journal of* Mathematics *and Computer Science Research*, 2013: 26-34.

[18] Ryan, J. V. "Classification and Clustering." (Academic Press Inc) 1977.

[19] Sneath, P., and Sokal, R. Numerical Taxonomy. W. H. Freeman Co., San Francisco, CA, 1973.

[20] Tarpey, T. (2007). Linear transformations and the k-means clustering algorithm. The American Statistician, 61, 34–40.

[21] Wilmink, F. W. & Uytterschaut, H. T. (1984). Cluster analysis, history, theory and applications. In G. N. van Vark & W. W. Howells (Eds.), Multivariate Statistical Methods in Physical Anthropology (pp. 135-175). Dordrecht, The Netherlands: D. Reidel Publishing Company.