PREVALENCE OF INFECTIOUS DISEASES IN KATSINA STATE: AN INSIGHT USING CLUSTERING APPROACH

Dauda Usman
Department of Mathematics and Computer Science,
Umaru Musa Yar'adua University, Katsina.
Katsina-Nigeria.
Email: dauusman@gmail.com

Ibrahim Lawal Kane
Department of Mathematics and Computer Science,
Umaru Musa Yar'adua University, Katsina.
Katsina-Nigeria.

***Abstract*—**Data mining is a convenient way of extracting patterns, which represents knowledge implicitly stored in large datasets and focuses on issues relating to their feasibility, usefulness, effectiveness and scalability. This paper presents a hierarchical clustering analysis of infectious disease data of Katsina State. The diseases with a similar degree of prevalence were identified. The result of the cluster formation shows that Malaria is more prevalent in the State, followed by Cholera and Typhoid fever as shown by the Single Linkage and Centroid methods. The Complete Linkage and Ward methods showed that Malaria is the most prevalent followed by Typhoid fever and Cholera in Funtua and Katsina zones, while in Daura zone Typhoid fever is more prevalent followed by Malaria and Cholera.. The Chi-square test for independence indicates that the number of clusters tends to vary from one zone to another. The study concludes that clustering methods is a suitable tool for assessing the level of infections of the disease.

Key words: Malaria, Cholera, Typhoid fever and Cluster analysis.

I. INTRODUCTION

The term *cluster analysis* first used by **[1]** encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories. A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, to develop taxonomies. In an other word cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Given the above, cluster analysis can be used to discover structures in data without providing an explanation/interpretation. Clustering is one of the widely used knowledge discovery techniques to reveal structures in a data set that can be extremely useful to the analyst **[2]**.

Diseases affecting humans are caused by infection. Such as leprosy, chicken pox and typhoid fever **[3]**. The aetiology of some of these diseases is induced by environmental factors.

Infection differs from other diseases in a number of aspects. The most important is

that it is caused by living microorganisms which can usually be identified, thus establishing the aetiology early in the illness. Many of these organisms, including all bacteria, are sensitive to antibiotics and most infections are potentially curable, unlike many non-infectious diseases which are degenerative and frequently become chronic. Communicability is another factor which differentiates infectious from non-infectious diseases. Transmission of pathogenic organisms to other people, directly or indirectly, may lead to an epidemic. Finally many infections are preventable by hygienic measures, by vaccines or by the judicious use of drugs (chemoprophylaxis) **[4]**. For these reasons, therefore, statisticians and social scientists use different scientific methods to analyze the cultural and behavioral aspects of the infectious diseases as well as their impact on families, communities and nations in general.

One of the most commonly used scientific methods is multivariate analysis. Multivariate analysis consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples. We will refer to the measurements as variables and to the individuals or objects as units (research units, sampling units, or experimental units) or observations. In practice, multivariate data sets are common, although they are not always analyzed as such.

But the exclusive use of univariate procedures with such data is no longer excusable, given the availability of multivariate techniques and inexpensive computing power to carry them out.

Historically, the bulk of applications of multivariate techniques have been in the behavioral and biological sciences. However, interest in multivariate methods has now spread to numerous other fields of investigation. For example, in education, chemistry, physics, geology, engineering, law, business, literature, religion, public broadcasting, nursing, mining, linguistics, biology, psychology, and many other fields **[5a]**.

[6] Opined 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 variable necessary to describe it.

A clustering technique (one of the multivariate approach), has been applied to a wide variety of research problems. **[7a]** provides an excellent summary of the many published studies reporting the results of cluster analysis. For example, in the field of medicine, clustering of disease, cures for diseases, or symptoms of diseases have led to very useful taxonomies. In the field of psychiatry, the correct diagnosis of cluster of symptoms such as paranoia, schizophrenia is essential for successful therapy

Cluster analysis seeks to partition a set of individuals into some form of natural groupings, if any. It is one tool of exploratory data analysis that attempt to

assess the interaction among patterns by organizing the patterns into groups or cluster, such that patterns within a cluster are more similar to each other than are pattern belonging to different clusters **[7b]**.

[8] Applied hierarchical clustering technique to partition the set of variables into groups, such that those that are similar with respect to HIV/AIDS infections were identified and two main clusters were observed, and 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 diseases affect mostly the working class aged from 15 to 39 years as grouped by complete linkage method.

[9] Analyzed data of infectious diseases of two senatorial zones of Katsina State using hierarchical clustering technique and partitioned the set of the diseases into groups such that the diseases with a similar degree of prevalent were identified, indicates the most two prevalent diseases in each zone. **[10]** Applied hierarchical clustering technique for the origins of the new influenza A(H1N1) virus and reported that the virus was the reassortment of at least two swine influenza viruses from North America (in light blue) and Eurasia (in dark blue).

II. MATERIALS AND METHODS

The method employed for the classification of the diseases according to three senatorial zones in Katsina State is hierarchical clustering techniques which include: Single Linkage Method, Complete Linkage Method, Centroid Method and Ward's Method. These methods are generally suitable for searching of natural clusters and they perform reasonably well when clusters are clearly separated **[11a]**. The four linkage methods were used, as this will help to prevent misleading results being accepted. However the differences in the linkage methods are due to differences in defining distance (similarity) between groups for each of the methods **[11b]**.

A. MEASURES OF PROXIMITY

Cluster analysis attempts to identify the observation vectors that are similar and group them into clusters, many techniques use an index of proximity between each pair of observations. A convenient measure of proximity is the distance between two observations. Since a distance increases as two units become further apart, the distance is actually a measure of dissimilarity.

A common distance function is the Euclidean distance between two vectors

$X = \left(x_1, x_2, ..., x_p\right)'$ and

$Y = \left(y_1, y_2, ..., y_p\right)'$, defined as:

$$d\left(x, y\right) = \sqrt{\left(x - y\right)'\left(x - y\right)} = \sqrt{\sum_{j=1}^{p}\left(x_j - y_j\right)^2}.$$

(3.1)

To adjust for differing variances and covariances among the p variables, we could use the statistical distance

$$d\left(x, y\right) = \sqrt{\left(x - y\right)' s^{-1}\left(x - y\right)}$$

(3.2)

where S is the sample covariance matrix. After the clusters are formed, S could be computed as the pooled within-cluster covariance matrix.

B. HIERARCHICAL ALGORITHMS (AGGLOMERATIVE TECHNIQUES)

The hierarchical attempt to find "good" clusters in the data using a computationally efficient technique. The method is also used quite frequently in practice, the algorithm consists of the following steps:

1. Construct the finest partition.
2. Compute the distance matrix D.
3. Find the two clusters with the closest distance.
4. Put those two clusters into one cluster.
5. Compute the distance between the new groups and obtain a reduced distance matrix D.

UNTIL all clusters are agglomerated into X. **[12]**.

C. SINGLE LINKAGE (NEAREST NEIGHBOR)

In the single linkage method, the distance between two clusters A and B is defined as the minimum distance between a point in A and a point in B:

$$D(A,B) = \min\{ d(y_i, y_j), \text{for } y_i \text{ in } A \text{ and } y_j \text{ in } B \}, \quad (3.3)$$

where $d(y_i, y_j)$ is the Euclidean distance in (3.1) or some other distance between the vectors y_i and y_j. This approach is also called the nearest neighbor method. At each step in the single linkage method, the distance (3.3) is found for every pair of clusters, and the two clusters with smallest distance are merged. The number of clusters is therefore reduced by 1. When two clusters are merged, the procedure is repeated for the next step: the distances between all pairs of clusters are calculated again, and the pair with the minimum distance is merged into a single cluster. **[5b]**.

D. COMPLETE LINKAGE (FARTHEST NEIGHBOR)

In the complete linkage approach, also called the farthest neighbor method, the distance between two clusters A and B is defined as the maximum distance between a point in A and a point in B:

$$D(A,B) = \max\{ d(y_i, y_j), \text{for } y_i \text{ in } A \text{ and } y_j \text{ in } B \}, \quad (3.4)$$

At each step, the distance (3.4) is found in every pair of clusters, and the two clusters with the smallest distance are merged. **[5c]**.

E. CENTROID METHOD

In the centroid method, the distance between two clusters A and B is defined as the Euclidean distance between the mean vectors (often called centroids) of the Two clusters:

$$D(A,B) = d(\bar{y}_A, \bar{y}_B), \quad (3.5)$$

where $\overline{y_A}$ and $\bar{y}_B$ are the mean vectors for the observation vectors in A and the observation vectors in B, respectively,

and $d\left(\bar{y}_{A,}\bar{y}_{B}\right)$ is defined in (3.1). We define $\bar{y}_{A}$ and $\bar{y}_{B}$ in the usual way, that is, $\bar{y}_{A}=\sum_{i=1}^{nA}\frac{yi}{nA}$. The two clusters with the smallest distance between centroids are merged at each step.

After two clusters A and B are joined, the centroid of the new cluster AB is given by the weighted average **[5d]** as:

$$\bar{y}_{AB}=\frac{n_A\bar{y}_A+n_B\bar{y}_B}{n_A+n_B}. \quad (3.6)$$

F. WARD'S METHOD

Ward's method, also called the incremental sum of squares method, uses the within cluster (squared) distances and the between-cluster (squared) distances. If AB is the cluster obtained by combining clusters A and B, then the sum of within-cluster distances (of the items from the cluster mean vectors) are:

$$SSE_A=\sum_{i=1}^{n_A}\left(y_i-\bar{y}_A\right)'\left(y_i-\bar{y}_A\right), \quad (3.7)$$

$$SSE_B=\sum_{i=1}^{n_B}\left(y_i-\bar{y}_B\right)'\left(y_i-\bar{y}_B\right), \quad (3.8)$$

$$SSE_{AB}=\sum_{i=1}^{n_{AB}}\left(y_i-\bar{y}_{AB}\right)'\left(y_i-\bar{y}_{AB}\right), \quad (3.9)$$

Where $\bar{y}_{AB}=\frac{\left(n_A\bar{y}_A+n_B\bar{y}_B\right)}{\left(n_A+n_B\right)}$, as in (3.6), and $n_A,n_B,$ and $n_{AB}=n_A+n_B$ are the numbers of points in A, B, and AB, respectively. Since these sums of distances are equivalent to within-cluster sums of squares, they are denoted by: SSE_A,SSE_B and SSE_{AB} **[5e]**.

Ward's method joins the two clusters A and B that minimize the increase in SSE, defined as

$$I_{AB}=SSE_{AB}-\left(SSE_A+SSE_B\right). \quad (3.10)$$

G. TEST FOR INDEPENDENCE

The hypothesis we wish to test for is whether the number of clusters formed by different methods varies from one zone to the other, using a test of independence. The procedure for the test is:

1. We set the hypothesis as:

H_1: Number of clusters formed by different methods vary from one zone to another.

H_0: Number of clusters formed by different methods do not vary from one zone to another.

2. We choose a level of significance α $=0.5$

3. The test statistic under the null hypothesis is $\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(0_{ij}-e_{ij}\right)^2}{e_{ij}}$ which is a distributed approximately as $X^2{}_{\alpha}\ (v)$ variant with $(r-1)(c-1)$ degrees of freedom O_{ij} is the observed frequency while e_{ij} is the expected frequency in the cell.

4. Determine the expected values under H_O in each cell which is the product of the marginal totals common to that cell divided by the total number of

observations and calculate the value of χ^2. Also determine the degrees of freedom *(r-1)(c-1)*.

5. The region of rejection is determined with the help of degree of freedom, the level of significance and from χ^2 tables.

6. Critical region: reject H_o if the calculated value is greater than the tabulated value at *α%* level of significance and degree of freedom (*r-1)(c-1*), otherwise fail to reject.

III. RESULTS AND DISCUSSIONS

A. RESULTS FROM FUNTUA ZONE

Data from Funtua zone was analyzed using all the four cluster formation methods and discussed in the table 1 below.

Table: 1 Summary of cluster formation analysis of Funtua zone data.

Cluster	Methods for cluster formation			
	Single Linkage	Complete Linkage	Centroid Method	Ward Method
1	5	5	5	5
2	5,6	5,6	5,6	5,6
3	5,6,4	5,6,4	5,6,4	5,6,4
4	5,6,4,3	5,6,4,3	5,6,4,3	5,6,4,3
5	5,6,4,3,8	5,6,4,3,8	5,6,4,3,8	5,6,4,3, 8

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table 1 presents the summary of the cluster formation analysis for Funtua zone data and indicates that Malaria is the most prevalent disease in Funtua zone, irrespective of the cluster formation method employed. It is closely followed by cholera, irrespective of the method employed.

Malaria and Cholera were followed by Typhoid fever, then Chickenpox and finally Measles. The five diseases from a single cluster of closely related diseases that are statistically significant.

B. RESULTS FROM DAURA ZONE

Data from Daura zone was analyzed using all the four cluster formation methods and discussed in the table 2 below.

Table: 2 Summary of cluster formation analysis of Daura zone data.

Cluster	Methods for cluster formation			
	Single Linkage	Complete Linkage	Centroid Method	Ward Method
1	5	4	5	4
2	5,6	4,5	5,6	4,5
3	5,6,4	4,5,6	5,6,4	4,5,6
4	5,6,4,8	4,5,6,8	5,6,4,8	4,5,6,8
5	5,6,4,8,3	4,5,6,8,3	5,6,4,8,3	4,5,6,8,3

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table 2 presents the summary of the cluster formation analysis for Daura zone data and indicates that Malaria is more prevalent using Single and Centroid methods and Typhoid fever using Complete and Ward Methods. They are closely followed by Cholera using Single and Centroid methods and Malaria using Complete and Ward methods.

They were also followed by Typhoid fever, Measles and Chicken Pox for Single and Centroid methods and Cholera, Measles and Chicken Pox for Complete and Ward methods.

C. RESULTS FROM KATSINA ZONE

Data from Katsina zone was analyzed using all the four cluster formation methods and discussed in the table 3 below.

Table: 3 Summary of cluster formation analysis of Katsina zone data.

Cluster	Methods for cluster formation			
	Single Linkage	Complet eLinkage	Centroid Method	Ward Method
1	5	5	5	5
2	5,6	5,4	5,6	5,4
3	5,6,4	5,4,6	5,6,4	5,4,6
4	5,6,4,3	5,4,6,3	5,6,4,3	5,4,6,3
5	5,6,4,3,8	5,4,6,3,8	5,6,4,3,8	5,4,6,3,8

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table 3 presents the summary of the cluster formation analysis for Katsina zone data and indicates that Malaria is most prevalent disease in Katsina zone irrespective of the cluster formation method employed. It is closely followed by Cholera using Single and Centroid methods and Typhoid fever using Complete and Ward methods.

They were also followed by Typhoid fever, Chicken Pox and Measles using Single and Centroid methods and Cholera, Chicken Pox and Measles.

D. RESULT FOR THE TEST OF INDEPENDENCE

Table 4: Test for Independence.

Table 4 presents the calculated chi-square

	Funtua	**Daura**	**Katsina**	**Total**
Single Linkage	5 (5)	5 (5)	5 (5)	15
Complete Linkage	5 (5)	5 (5)	5 (5)	15
Centroid Method	5 (5)	5 (5)	5 (5)	15
Ward Method	5 (5)	5 (5)	5 (5)	15
	20	20	20	60

Table 4 presents the calculated chi-square as follows:

$$\chi^2 = (5-5)^2/5 + \ldots\ldots (5-5)^2/5 = 0$$

The tabulated chi-square is as follows:

$$\chi^2_{\alpha(r-1)(c-1)} = \chi^2_{0.05,6} = 1.64$$

Decision: we fail to reject H_0 and conclude that the number of cluster formation tend to vary from one zone to another when different methods are employed.

IV. CONCLUSION

The results of the study show that Malaria is more prevalent in all the three senatorial zones followed by Cholera and Typhoid fever. The prevalence of these diseases may be due to the nature of some areas in the state, where we have rivers, ponds where the causative agent (mosquitoes) can breed easily. Some of these diseases may also come from contaminated food, water and even fruits such as mangoes, cashew fruits, pawpaw etc. usually caused by housefly and the tsetse fly. Malaria being the most prevalent disease followed by Cholera and Typhoid fever in all the three zones in Katsina State, therefore there is the need for government or authority concerned to put in place sound

programs for the eradication of such diseases and provides welfare services such as drainage system, environmental protection and good pipe borne water.

REFERENCES

[1] Tryon, R. C. (1939). *Cluster analysis.* New York: McGraw-Hill.

[2] Han, J., Kamber, M. Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.

[3] Bloom, A. (1963), *Toohey's Medicine for Nurses.* Whittington London.

[4] Davidson, S. (2006), *Principle and Practice of Medicine, 20th Edition.* Edinburge, New York.

[5a-e] Rencher, A. C. (2002), *Method of Multivariate Analysis 2nd edition.* John Wiley & Sons Inc, Third Avenue, New York.

[6] Morrisson, D.F. (1990), *Multivariate Statistical Methods*. McGraw-Hill Book Company, New York.

[7a,b] Hartigan, J.A. (1972), Direct clustering of a data matrix. *Journal of the American Statistical Association*. 67, 123 – 129.

[8] Gulumbe, S. U., Bakar, A.B. and Dikko, H.G. (2008), Classification of some HIV/AIDS Variables, a multivariate approach. *Research Journal of Science* 15, 24 – 30.

[9] Dauda, U., Gulumbe, S.U., Yakubu, M. and Ibrahim, L.K. (2011), Monetering of Infectious Diseases in Katsina and Daura Zones, a Hierarchical Cluster Analysis. *Nigerian Journal of Basic and Applied Sciences* 19(1), 31-42

[10] Solovyov, A., Palacious, G., Briese, T., Lipkin, W.I. and Rabadan, R. (2009), Cluster Analysis of the Origins of the New Influenza A (H1N1) Virus. *Eurosurveillance Journal 14, 38 – 41.*

[11a,b] Everitt, B.S. (1974), *Cluster Analysis*. Heinemann Educational Books Limited U.K.

[12] Hardle, W. and Simar, L. (2007), *Applied Multivariate Statistical Analysis.* Spring Berlin Heidelberg, New York.