Clustering the Districts of Sri Lanka Using Hierarchical Methods: An Application to a Computer Literacy Study

H T Tharanganie[1*]

[1] Department of Statistics, University of Colombo, Colombo 03, Sri Lanka.

Abstract

This paper is based on a study to cluster the 19 districts of Sri Lanka on the results of the 31 variables which influence the computer literacy of households. These variables were selected from the Household Computer Literacy Survey of Sri Lanka-2006/2007 conducted by the Department of Census and Statistics in years 2006 and 2007 to assess the computer literacy of household population in the ages of 5 to 69 years and e-readiness of households. In order to achieve this classification, first the descriptive analysis and then the advanced analysis were accomplished. In the descriptive analysis, multivariate data were presented graphically using star plots. Those plots clearly indicate Nuwara-Eliya and Monaragala districts are outliers since they have been represented by out of star-shaped plots. From the graphical display, distinguishable groupings cannot be identified. Therefore using five hierarchical cluster methods (Ward's, Complete Linkage, Single Linkage, Average Linkage and Centroid Linkage), an advanced analysis was performed to identify well-separated clusters.

In all five methods, Colombo district remains a cluster by itself clearly separating from other districts. Ward's method classifies districts into three clusters including Nuwara-Eliya and Monaragala districts as a one cluster and this would appear to be reasonable since these two are identified as outliers in the graphical representation. Nevertheless other four methods classify districts into two clusters and in them, Monaragala district is merged at last at a higher distance and Nuwara-Eliya district is merged just before the Monaragala district. This further reveals that these two districts are more similar to each other than to the other districts. Hence, three clusters of districts are identified and the cluster structure is as follows. Colombo district remains a cluster by itself, both Nuwara-Eliya and Monaragala districts perform the second cluster while the rest of the 16 districts fall into the third cluster. Afterwards to access the suitability of these clusters, the mean profiles of three clusters against the 31 variables were plotted to evaluate whether there is any redundancy among the clusters. Since mean profiles of three clusters are not piecewise parallel reveal that the three clusters are not equivalent and they do not carry any redundant information.

Introduction

Household Computer Literacy Survey of Sri Lanka was conducted by the Department of Census and Statistics in years 2006/2007 to assess the computer literacy of household population and e-readiness of households. This survey had been covered the 19 districts of the country other than Northern Province and Trincomalee district in the Eastern Province. Clustering these 19 districts according to the performance of the selected 31 variables which influence the computer literacy of households will be an objective of this study. In order to achieve this objective, a graphical display and a cluster analysis were accomplished with the assistance of STATISTICA and SAS softwares.

Methodology

In this study, star plots were used as a graphical display. Then, five hierarchical methods (Ward's, Complete Linkage, Single Linkage, Average Linkage and Centroid Linkage) were used to identify the clusters. Finally, the mean profiles of clusters were plotted to evaluate whether there is any redundancy among the clusters.

Results

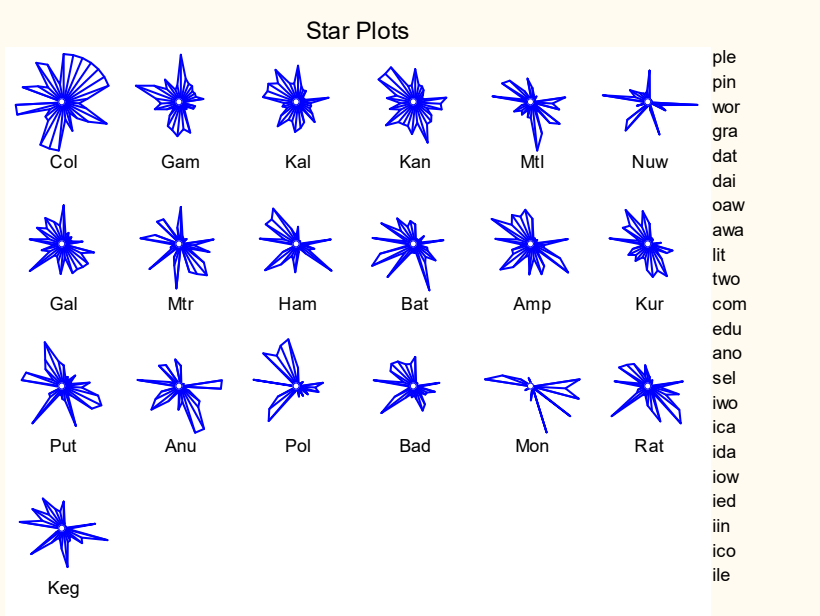

Figure 1: Star Plots

It can be seen that Nuwara-Eliya (Nuw) and Monaragala (Mon) districts have been represented by out of star-shaped plots. Since these two are considerably different from others, they are said to be outliers. Except these two districts, stars representing other districts in the above plot are virtually indistinguishable into well-separated clusters. From the graphical aids in clustering, clear clusters cannot be identified and an advanced analysis is needed to perform. The data in this study were standardized by including a

STANDARD option in SAS statistical software before performing the cluster analysis. Thus *Z* scores were used by the CLUSTER procedure[1].

Ward's cluster method - It can be seen from the dendrogram of Ward's method that the data fall into two distinct clusters. First including Colombo (Col) district and second including rest of the 18 districts. Further a cluster structure with three clusters can also be identified where the first including Colombo district, while the second including both Nuwara-Eliya and Monaragala districts and finally the third including the rest of the 16 districts. To determine the appropriate number of clusters, the Hotelling's pseudo T^2 statistic (PST2) would be concentrated[1]. PST2 can be used to help determine whether the two clusters combined should have been combined. If PST2 is large, the two clusters should not be combined; but if PST2 is small, then the two clusters can safely be combined. It is also advisable to ignore the PST2 values greater than 20% of the number of data points, which is 20%*19 = 3.8 for these data. Hence, only the PST2 values which are lesser than 4 are considered in identifying the number of clusters. When the program reduces from 4 clusters to 3 clusters, the PST2 is 2.5. When one goes from 3 clusters to 2 clusters, the PST2 is 4.2, which is kind of large compared to 2.5. Thus an examination of the PST2 values leads one to believe that the appropriate number of clusters is three.

Complete Linkage cluster method - Dendrogram of Complete Linkage method illustrates two clear separate clusters. The first including Colombo district and Gampaha (Gam) district while the second including rest of the 17 districts. However, Colombo and Gampaha districts do not merge until the distance between furthest neighbors has increased substantially. Further in the second cluster, Monaragala district is merged with the second cluster at a higher distance level and Nuwara-Eliya district is merged with the second cluster just before the Monaragala district.

Single Linkage, Average Linkage and Centroid Linkage cluster methods - With respect to the dendrograms of above three methods, two clear clusters are evident as Colombo district remains a cluster by itself and the rest of the 17 districts in the second cluster where the Monaragala district is merged at last at a higher distance and Nuwara-Eliya district is merged to the second cluster just before the Monaragala district.

Discussion

In all five hierarchical cluster methods, Colombo district remains a cluster by itself clearly separating from other districts. In four hierarchical cluster methods except Ward's

method, Monaragala district is merged at last at a higher distance and Nuwara-Eliya district is merged just before the Monaragala district in the second cluster. This reveals that these two districts are more similar to each other than to the other districts. This is confirmed by Ward's method as these two districts are joined and perform a cluster in a cluster structure of 3 clusters. In Complete Linkage cluster method, since merging of Colombo and Gampaha districts does not occur at a small distance between furthest neighbors and this merging is only seen in this method, a cluster with both Colombo and Gampaha districts can be ignorable. However, the dendrogram of Ward's method seems to be distinct from others. It is known that in general, this method is regarded as very efficient as it is based on minimizing the 'loss of information' from joining two groups. Thus an examination of the PST2 values of the Ward's method, leads one to believe that the appropriate number of clusters is three. In this method, both Nuwara-Eliya and Monaragala districts perform a one cluster and this would appear to be reasonable since these two are identified as outliers in the graphical representation star plots. Summarizing the information above, it can be concluded that the number of clusters in these data is three and they are as follows. Colombo district remains a cluster by itself, both Nuwara-Eliya and Monaragala districts perform the second cluster while the rest of the sixteen districts fall into the third cluster.

In order to perform a Multivariate Analysis of Variance (MANOVA) to access the suitability of these clusters, the data to be Multivariate Normal distributed and it is checked using the Chi-square plot[2]. But this can be constructed only if both the size of the sample (n) and the difference between the size of the sample and the number of variables in the sample (n-p) are greater than 25. Since in this study, only the 19 observations (n=19) with 31 variables (p=31) are considered, the Chi-square plot is unable to drawn to check the multivariate normality. Hence, MANOVA cannot be used to access the suitability of these clusters, instead a plot of the mean profiles of three clusters is allowed to evaluate whether there is any redundancy among the clusters.

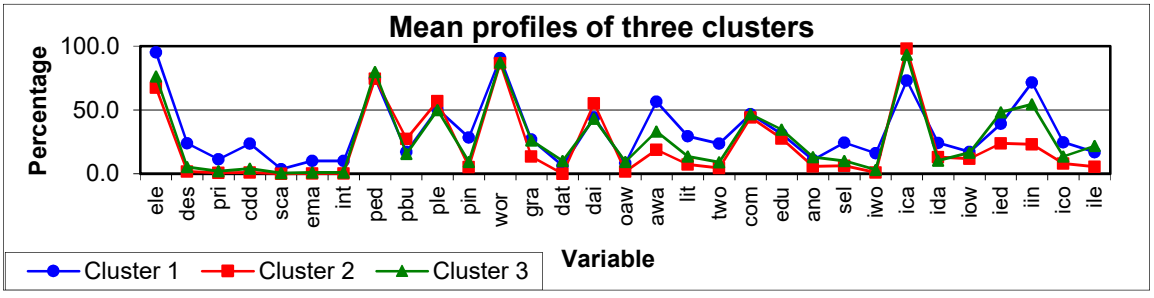

Figure 2: Plot of the mean profiles of three clusters

As shown in the above figure, since mean profiles of three clusters are not piecewise parallel suggest that the three clusters are not equivalent. It is also evident that the mean profiles of only one variable (which is ‘com’) are nearly coincided whereas in other variables, mean profiles of three clusters show significant differences. Therefore, it ensures that the three clusters do not carry any redundant information, since any one cluster is different the other at least from a single variable.

References

[1] Johnson, D. E., (1998). Applied Multivariate Methods for Data Analysts (2nd ed.). Pacific Grove: Brooks/Cole Publishing Company.

[2] Johnson, R. A., & Wichern, D. W. (2002). Applied Multivariate Statistical Analysis (5th ed.). London: Prentice Hall.