



511/E1

Clustering the districts of Sri Lanka using hierarchical methods: An application to a computer literacy study

H T Tharanganie*

Department of Statistics, University of Colombo, Colombo 3.

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 that 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.