IJIRR

*International Journal of Information Research and Review*
***Vol. 04, Issue, 05, pp.4128-4132, May, 2017***

## RESEARCH ARTICLE

# A MULTIVARIATE APPROACH TO CLASSIFY THE DISTRICTS OF SRI LANKA BASED ON THE COST OF LIVING

***Sebastian Reyalt Gnanapragasam**

Lecturer, Department of Mathematics and Computer Science, the Open University of Sri Lanka

**ARTICLE INFO**

***Article History:***

Received 04th February, 2017
Received in revised form
12th March, 2017
Accepted 01st April, 2017
Published online 30th May, 2017

***Keywords:***

Classification,
Cluster Analysis,
Cost of living,
Districts of Sri Lanka,
Principle Component Analysis

**ABSTRACT**

Cost of Living (CoL) differs among 25 districts in Sri Lanka. The objective of the study is to classify the districts of Sri Lanka based on the CoL. To achieve this, districts were ranked and were clustered. This will support the relevant officials and decision makers in the country when district-wise infra-structure development planning taking place. In the data, two types of variables were considered, one as food items and the other as non-food items. Descriptive statistics such as variances and correlations were used to observe the variance of variables and to check the associations among variables respectively. Principle Component Analysis was employed to rank the districts. Cluster Analysis was used to group the districts. This study revealed that the average expenditure at a district was 41,444 LKR per month in which only about 37.76% of total average expenditure was for food items. Also the results indicated that, CoL mainly depends on the non- food items and the expenditure on food items had less variation among districts. Thus it can be stated that the non-food items were the most influential factor in terms of CoL. It can be concluded that, the most expensive three districts were Colombo, Gampaha and Kalutara while Moneragala and Mullaitivu were the two least expensive districts. Further, it concluded that, Colombo was isolated from all other districts while only Gampaha and Kalutara formed a single group when three and four cluster classifications were considered.

## INTRODUCTION

Sri Lanka is an island in south Asia with 65,610 km$^2$ land area and slightly higher than twenty million people are living in the country. It has twenty five administrative divisions as districts in nine provinces. The cost of living differs among those districts in Sri Lanka. Basically the cost of living is the amount of money needed to maintain a certain level of living including basic expenses such as housing, food, education and health. Usually it is used to compare the expenses to live among cities or states or countries. It is identified that the cost of living in Sri Lanka can be categorized into several main sections such as markets, transportation, utilities, health, rent, taxes, clothing, and salary. From the information available in the island, it is clearly noted the variations in cost of living between several parts of the country. To measure the cost of living in Sri Lanka, at earliest the compilation of consumer price index (CPI) commenced in the early 1940s with the computation of the Colombo working class index (CWCI) and the estate labor index number (ELIN). Then the Colombo consumer's price index (CCPI) which replaced the CWCI in 1949/50. Next the greater Colombo consumers' price index (GCPI) was introduced in 1989. Later, the Colombo district consumer price index (CDCPI) was introduced by the Central Bank of Sri Lanka in 1998. Economists urged the need of a new CPI to eliminate many of its current issues and limitations such as exclusion of rent and expenses for education in CPI. The economic analysts pointed out that the new CPI should be based on a specially designed household expenditure survey with assigned weights for CPI. This study may be an initiative of creating the new index as the data used in this study include all twenty five districts by considering the expenses for house rent and education etc.

The main aim of this study is classifying the districts in Sri Lanka based on the cost of living. To attain this aim, the following objectives are achieved:

- To rank the districts based on cost of living in Sri Lanka
- To group the districts based on expenditure of essentials non-food items in Sri Lanka

This study will help the ones who want to decide the living district based on their income. It can be decided whether the expenses are manageable in that particular district based on their salary and accordingly the most suitable district can be chosen to live.

****Corresponding author: Sebastian Reyalt Gnanapragasam,***
*Lecturer, Department of Mathematics and Computer Science, the Open University of Sri Lanka.*

Further, it will support the relevant officials and decision makers in the country when district-wise infra-structure development planning taking place.

## MATERIALS AND METHODS

The data for this study is obtained from the latest “Household Income and Expenditure Survey Report 2012/13” and the report was released by the Department of Census and Statistics of Sri Lanka in March 2015. Two types of variables are considered in that survey one as food items and the other as non-food items. For the use of the software conveniently, fourteen food items are named as Y’s whereas thirteen non-food items are named as X’s as follows:

| Food Items |
|---|
| **Y1**= Cereals |
| **Y2**= Coconuts |
| **Y3**= Prepared Food |
| **Y4**= Condiments |
| **Y5**= Pulses |
| **Y6**= Milk& Milk Food |
| **Y7**= Vegetables |
| **Y8**= Fat & Oil |
| **Y9**= Meat |
| **Y10**= Sugar, Jungery & Treacle |
| **Y11**= Fish |
| **Y12**= Fruits |
| **Y13**= Dried Fish |
| **Y14**= Other Food Items |

| Non-Food Items |
|---|
| **X1**= Housing |
| **X2**= Fuel & Light |
| **X3**= Transport |
| **X4**= Communication |
| **X5**= Education |
| **X6**= Household Durable Goods |
| **X7**= Personal Care & Health Expenses |
| **X8**= Cultural Activities & Entertainments |
| **X9**= Household Non-durable Goods & Household services |
| **X10**= Clothing Textiles & Foot Wear |
| **X11**= Other Missellaneous Expenses |
| **X12**= Other Adhoc (rarely) Expenses |
| **X13**= Liquor, Drugs & Tobacco |

### Methods employed for preliminary analysis

Prior to classify the districts of Sri Lanka, descriptive statistics are used to describe the variables and to observe the associations among those variables on both types.

At preliminary level, mean is calculated to obtain the average expenditure on particular variable and variance-correlation matrices are obtained to check the associations among variables in both types separately.

### Principle Component Analysis (PCA)

PCA is applied to rank the districts based on the cost of living. Particularly in PCA, the following techniques are carried out:

Eigen analysis is used to extract the number of components. In general, if Eigen values are greater than one then those principal components (PCs) can be extracted to represent the new orthogonal system. Also it is better to make sure at least 75% of variability in original system is absorbed by the selected PCs.

Principle component coefficients are considered to estimate the values of principle component of each district and these values are used to rank the districts based on the cost of living.

### Cluster Analysis (CA)

CA is applied to group the districts based on the expenditure of essential items. Especially in CA, the following techniques are carried out:

**Determination of number of clusters:** Determining number of clusters is fundamental problem in CA. In this study, the following two techniques are employed for this purpose.

- Rule of thumb: number of clusters $\approx\sqrt{\left(\frac{n}{2}\right)}$ , where $n$ is the total number of observations.
- Elbow shape in Scree plot: In the plot of percentage of variance and number of clusters, the number of clusters is chosen at the point where an “elbow” shape can be first detected.

### Squared Euclidean distance

$d_{ij}^2 = \sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2$ , is taken as the distance measure, where xj xj $x_j$ and $x_j$ are the two cases being compared on the $k^{th}$ variable, and $P$ is the total number of variables extracted in the analysis. Linkage methods are used to measure the distance between two clusters. In this study, the following linkage methods are taken into account.

- **Average linkage method:** In the average linkage method, the distance between two clusters is the mean distance between an observation in one cluster and an observation in the other cluster.
- **Single linkage method:** In single linkage method, the distance between two clusters is the minimum distance between an observation in one cluster and an observation in the other cluster.
- **Centroid linkage method:** In the centroid linkage method, the distance between two clusters is the distance between the cluster centroids or means.
- **Complete linkage method:** In the complete linkage method, the distance between two clusters is the maximum distance between an observation in one cluster and an observation in the other cluster.
- **Ward’s linkage method:** In Ward's linkage method, the distance between two clusters is the sum of squared deviations from points to centroids.
- **Dendogram:** Finally dendrograms are obtained to identify and separate the groups.

### Preliminary Analysis

Figure 1 shows the district wise average monthly household expenditure.

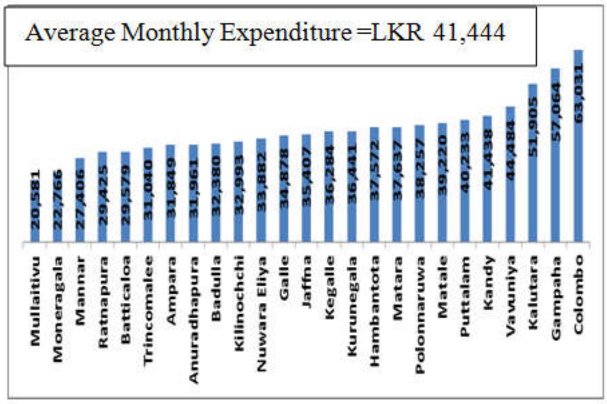

**Figure 1. Average monthly household expenditure**

As per Figure 1, it can be clearly seen on the average that, Mullaitivu is the least expensive district while Colombo is the most expensive district. Further it reveals that the mean of the expenditure at a district is LKR 41, 444.

The average monthly expenses of the districts Vavuniya, Kalutara, Gampaha and Colombo are above the mean expenditure and all other districts are below the mean expenditure except the Kandy district which is very closer to the mean expenditure.

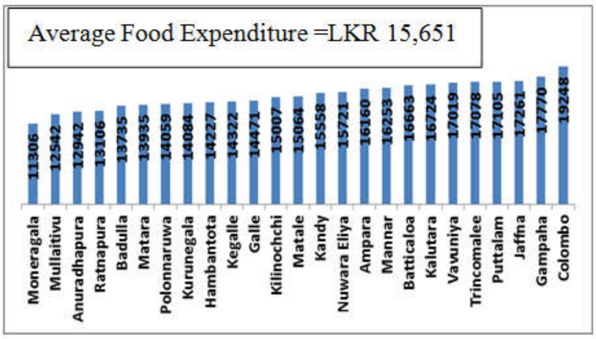

**Figure 2. Average monthly food expenditure**

According to the information in Figure 2, it can be stated that the variation among the districts is less when compare with that of in Figure 1 and ranges from LKR 11,306 to LKR 19,248. Based on the average expenses on food items, Monaragala is the least expensive district where as Colombo district is the most expensive on food item too. Further, it is noted that, the average monthly expenditure on food items at a district is LKR 15,651 and which is only about 37.76% of household expenditure.

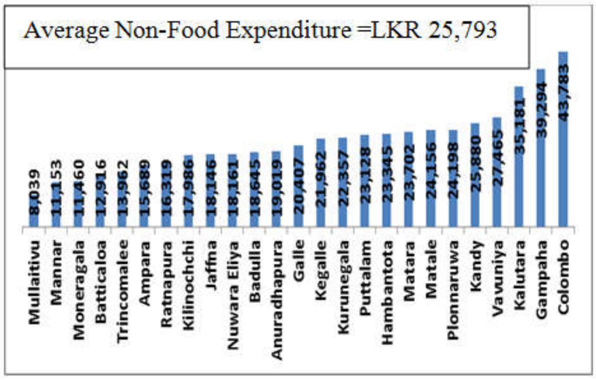

**Figure 3. Average monthly non-food expenditure**

**Table 1. Descriptive statistics of food items**

| Variable | Mean | Variance |
|---|---|---|
| Cereals | 2759 | 207311 |
| Prepared Food | 1486 | 422297 |
| Pulses | 517 | 16409 |
| Vegetables | 1216 | 26487 |
| Meat | 686 | 105816 |
| Fish | 1542 | 452994 |
| Dried Fish | 515 | 76602 |
| Coconuts | 968 | 19082 |
| Condiments | 1411 | 83229 |
| Fat & Oil | 413 | 13335 |
| Fruits | 427 | 15644 |
| Milk & Milk Food | 1270 | 94477 |
| Other Food Items | 1539 | 56066 |
| Sugar, Jungery& Treacle | 506 | 9633 |

From Figure 3 it can be clearly seen a big variation among the districts on the expenditure on non-food items on the average and the range is from LKR 8,039 to LKR 43,783. Also it is almost similar pattern like in average monthly household expenditure in Figure 1. Thus, it reveals that, monthly household expenditure depends on non- food items rather food- items. Further it is noted that, the mean of the expenditure on non-food items at a district is LKR 25, 793 and which is about 62.24 % of household expenditure. In this case too, like in all items expenditure in Figure 1, Mullaitivu is the least expensive district while Colombo is the most expensive district on the expense on non-food items. As per the statistics appear in Table 1, the people of Sri Lanka mainly spend for Cereals whereas they spend less for Fat & Oil and Fruit items. Here it is importantly noted that, there is significant differences in variances. It leads to carry out the multivariate analysis. According to the statistics in Table 2, it can be stated that, Sri Lankan people mainly spend for housing as non- food item next to all the other miscellaneous expenses. Meantime, they spend less for Cultural Activities & Entertain. Further it seriously noted that, the variances among the non-food items are significantly differentiate and hence it recommends employing multivariate analysis.

**Table 2. Descriptive statistics of non-food items**

| Variable | Mean | Variance |
|---|---|---|
| Housing | 3440 | 4873092 |
| Fuel & Light | 1551 | 340866 |
| Transport | 2753 | 1499960 |
| Communication | 731 | 97051 |
| Education | 1090 | 396089 |
| Liquor, Drugs & Tobacco | 685 | 63278 |
| Personal Care & Health Expenses | 1736 | 685054 |
| Cultural Activities & Entertain | 408 | 78610 |
| Household Durable Goods | 984 | 218046 |
| Clothing Textiles & Foot Wear | 1129 | 35466 |
| Other Miscellaneous Expenses | 3754 | 4269746 |
| Other Adhoc (rarely) Expenses | 2770 | 2457558 |
| Household Non-durable Goods & Household services | 424 | 25881 |

Note that, the red cells in Table 3 indicate that the relevant variables have no correlation whereas the green cells indicate that the relevant pair wise variables are strongly correlated. From Table 3, it can be observed that, very few variables in food items correlated with some of the other variables. However, most of the variables in food items have no strong pair wise correlation among other variables. Hence, these variables in food items cannot be considered further for multivariate techniques such as PCA and CA.

**Table 3. Correlation among the food items**

| | Y 1 | Y 2 | Y 3 | Y 4 | Y 5 | Y 6 | Y 7 | Y 8 | Y 9 | Y 10 | Y 11 | Y 12 | Y 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Y 2 | 🟥 | | | | | | | | | | | | |
| Y 3 | 🟥 | 🟥 | | | | | | | | | | | |
| Y 4 | 🟥 | 🟥 | | | | | | | | | | | |
| Y 5 | 🟥 | 🟩 | 🟩 | 🟥 | | | | | | | | | |
| Y 6 | 🟥 | 🟥 | 🟩 | 🟩 | | | | | | | | | |
| Y 7 | 🟩 | 🟥 | 🟩 | 🟩 | 🟥 | | | | | | | | |
| Y 8 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | | | | | | |
| Y 9 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | | | | | | |
| Y 10 | 🟥 | 🟩 | 🟩 | 🟥 | 🟥 | 🟥 | 🟩 | 🟥 | 🟥 | | | | |
| Y 11 | 🟩 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟩 | 🟥 | 🟥 | 🟥 | | | |
| Y 12 | 🟩 | 🟥 | 🟥 | 🟩 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | | | |
| Y 13 | 🟥 | 🟩 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟩 | 🟥 | 🟥 | |
| Y14 | 🟥 | 🟩 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟩 | 🟥 | 🟥 | 🟩 |

Note that, the green cells in Table 4 indicate that the relevant pair wise variables are strongly correlated on the other hand the red cells indicate that the relevant variables have no correlation. It is very clear from Table 4 that, very few of the pair wise variables in non-food items are not significantly correlated. On the other hand, most of the pairs of the variables in non-food items are significantly correlated. Hence, multivariate analyzing techniques are applicable for the variables in non-food items. Therefore for further analysis, only the variables in non-food items are taken.

**Table 4. Correlation among the non-food items**

| | X 1 | X 2 | X 3 | X 4 | X 5 | X 6 | X 7 | X 8 | X 9 | X 10 | X 11 | X 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X2 | 🟩 | | | | | | | | | | | |
| X3 | 🟩 | 🟩 | | | | | | | | | | |
| X4 | 🟩 | 🟩 | 🟩 | | | | | | | | | |
| X5 | 🟩 | 🟩 | 🟩 | 🟩 | | | | | | | | |
| X6 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | | | | | | |
| X7 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | | | | | |
| X8 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | | | | |
| X9 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | | | |
| X 10 | 🟩 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟩 | 🟥 | 🟥 | | | |
| X 11 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | |
| X 12 | 🟩 | 🟥 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | 🟥 | 🟩 | 🟩 | |
| X 13 | 🟩 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |

## Ranking the districts of Sri Lanka

In this section, the districts of Sri Lanka are ranked based on the cost of living.

**Table 5. Results of the Eigen analysis of PCA**

| Eigen value | Proportion | Cumulative proportion |
|---|---|---|
| 7.9 | 0.61 | 0.61 |
| 1.8 | 0.14 | 0.75 |
| 0.98 | 0.08 | 0.82 |
| 0.61 | 0.05 | 0.87 |
| 0.46 | 0.04 | 0.91 |
| 0.38 | 0.03 | 0.94 |
| 0.25 | 0.02 | 0.98 |
| 0.11 | 0.01 | 0.98 |
| 0.10 | 0.01 | 0.99 |
| 0.05 | 0.004 | 0.997 |
| 0.02 | 0.002 | 0.999 |
| 0.01 | 0.001 | 1.00 |

According to the results appear in Table 5, the first two Eigen values can be taken, as of those only the values are more than one. Further, it is clear from the scree plot of non-food items in Figure 4 that, the first elbow shape appears at the second component number.

Thus, first two principal components (PCs) can be taken to express original system in terms of orthogonal system. Moreover, from the cumulative proportion in Table 5, it can be stated that, the first two PCs consequently are able to acquire 75% of the variability of the original system. Nevertheless, to assign the rank of each of the districts in Sri Lanka, only the first PC is considered and the relevant coefficients of first PC are taken to rank the districts. Accordingly the ranked districts are ordered and listed in Table 8 under conclusion part.

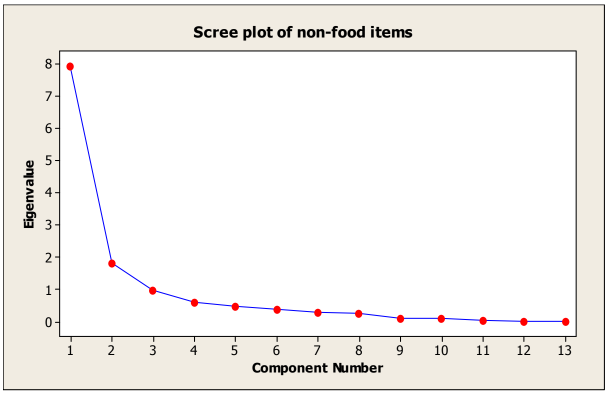

**Figure 4. Scree plot of non-food items**

## Clustering the districts of Sri Lanka

In this section, the districts are clustered as homogeneous groups based on the cost of living. The number of clusters is first calculated by the rule of thumb method. Since twenty five districts ($n$=25) are in Sri Lanka, number of clusters equals 3.54 and which falls in between 3 and 4. Thus, both cases are considered when clustering the districts.

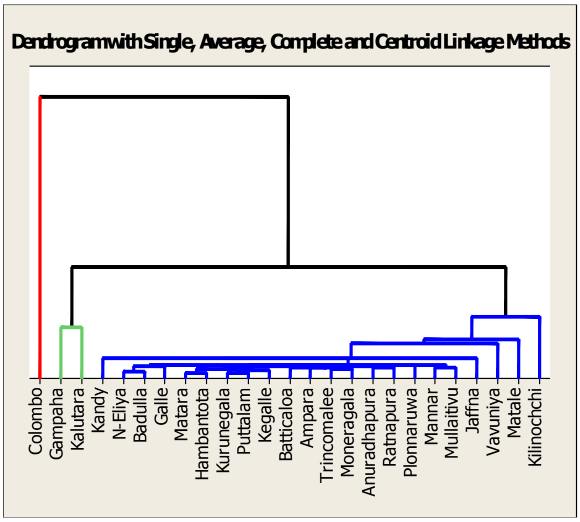

**Figure 5. Dendrogram with Single, Average, Complete and Centroid methods for three clusters**

Both tables Table 6 and Table 7 are summaries from the output of dendrograms, as appear in Figure 5 and Figure 6, obtained using all different five linkage methods.

**Table 6. Three cluster classifications**

| Method | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| Centroid | 1 | 2,3 | Others-22 districts |
| Average | 1 | 2,3 | Others-22 districts |
| Complete | 1 | 2,3 | Others-22 districts |
| Single | 1 | 2,3 | Others-22 districts |
| Ward's | 1,2,3 | 4,5,8,9,12,18,19,21,25 | Others-12 districts |

As per three cluster classifications appear in Table 6, except Ward's method all other method provide similar groupings. Since the majority four of all five methods suggest the same groupings, the districts can be classified into three clusters according to the first four linkage methods in Table 6 and the groupings are reported in conclusion part.

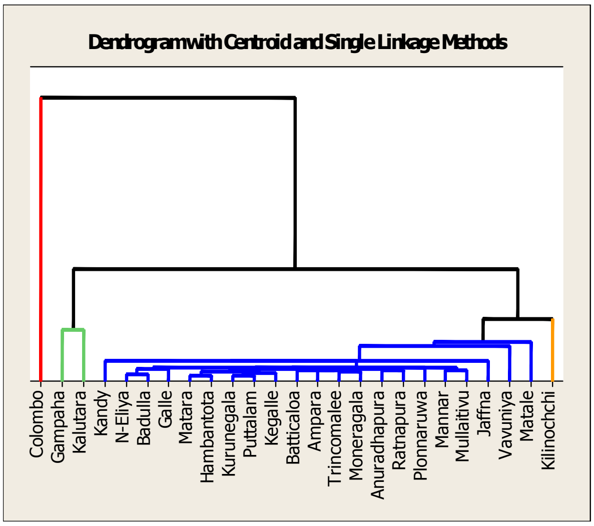

**Figure 6. Dendrogram with Single and Centroid methods for four clusters**

**Table 7. Four cluster classifications**

| Method | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| Centroid | 1 | 2,3 | Others-21 districts | 14 |
| Average | 1 | 2,3 | Others-20 districts | 5,14 |
| Complete | 1 | 2,3 | 4,6,7,8,9,12,18,19,20,21,22,24,25 | 5,10,11,13,14,15,16,17,23 |
| Single | 1 | 2,3 | Others-21 districts | 14 |
| Ward's | 1 | 2,3 | 4,5,8,9,12,14,18,19,21,25 | 6,7,10,11,13,15,16,17,20,22,23,24 |

**Table 8. Ranks of districts based on cost of living**

| Rank | District | Rank | District |
|---|---|---|---|
| 1 | Colombo | 14 | Anuradhapura |
| 2 | Gampaha | 15 | Galle |
| 3 | Kalutara | 16 | Badulla |
| 4 | Vavuniya | 17 | Nuwara Eliya |
| 5 | Kandy | 18 | Ampara |
| 6 | Matale | 19 | Trincomalee |
| 7 | Polannaruwa | 20 | Ratnapura |
| 8 | Matara | 21 | Batticaloa |
| 9 | Puttalam | 22 | Kilinochcchi |
| 10 | Hambantota | 23 | Mannar |
| 11 | Jaffna | 24 | Moneragala |
| 12 | Kurunegala | 25 | Mullaitivu |
| 13 | Kegalle | | |

According to four cluster classifications in Table 7, majority of five methods, Centroid and Single linkage methods suggest similar groupings. Hence, the districts can be classified into four clusters and the groupings are reported in conclusion part.

### Conclusion

Since more variations among districts can be seen on the expenditure on non-food items than food items, it can be claimed that, the cost of living mainly depends on non- food items. Moreover, on the average, nearly two third of total cost of living is for non- food items. Thus it can be concluded that the non-food items are more influence factors in terms of cost of living in Sri Lanka. Based on the cost of living, the districts of Sri Lanka are ranked and which are ordered and listed in Table 8. Also it can be concluded that, the most expensive three districts are Colombo, Gampaha and Kalutara while Mullaitivu and Moneragala are the two least expensive districts in Sri Lanka. If four clusters are needed, then four cluster classifications of districts based on cost of living are grouped as follows:

**Cluster 1**: Colombo

**Cluster 2**: Gampaha and Kalutara

**Cluster 3**: Kilinochchi

**Cluster 4**: Other 21 Districts

However, when three clusters are needed, then there are no changes in first two clusters but last two clusters can be merged together and accordingly three cluster classifications of districts based on cost of living are grouped as follows:

**Cluster 1**: Colombo

**Cluster 2**: Gampaha and Kalutara

**Cluster 3**: Other 22 Districts

Further, it can be concluded that, Colombo district is isolated from all other districts while only Gampaha and Kalutara form a single group when three and four cluster classifications are considered.

## REFERENCES

Financial Times, 2007. Colombo COL index outdated, Volume 42- No 11 (ISSN: 1391-0531). Retrieved from: http://www.sundaytimes.lk/070812/FinancialTimes/ft345.html

Household Income and Expenditure Survey Report 2012/13 2015. Department of Census and Statistics, Ministry of Policy Planning Economic Affairs, Child Youth and Cultural Affairs, Sri Lanka. Retrieved from: http://www.statistics.gov.lk/HIES/HIES2012_13FinalReport.pdf

Korale. RMB, 2001. "The Problem of Measuring Cost of Living in Sri Lanka", Macroeconomic Policy Series, IPS Publication. Retrieved from: http://www.ips.lk/index.php/35-pub-series/35-pub-series/929-the-problems-of-measuring-cost-of-living-in-sri-lanka

*******