**Tropical Agricultural Research Vol. 27 (1): 95 – 102 (2015)**
**Short Communication**

# Adjusting Variables in Constructing Composite Indices by Using Principal Component Analysis: Illustrated By Colombo District Data

J.S.N.P. Dharmawardena*, R.O. Thattil[1] and S. Samita[2]

Postgraduate Institute of Agriculture
University of Peradeniya
Sri Lanka

**ABSTRACT:** Constructing composite indices, variables have to be reduced and principal component analysis (PCA) is widely used for this purpose. Although, PCA is conducted after standardizing variables to overcome unit and value dependency problems, variables lose their inherent variability. To address that issue, two options were tested. First option was transforming variables by dividing their means, resulting new means and variances becoming one and square of coefficient of variance ($CV^2$) respectively. Second option was making meaningful adjustment to original variables to convert them as unitless. Grama Niladhari (GN) division level data on thirteen variables in Colombo district were used and second option was successful illustrating contribution of first two PCs to total variability by 96.34%. However, in the conventional method, 8 PCs were needed to reach that proportion. Expressing some variables on a per household basis and dividing GN density by total density were two adjustments made and that resulted in meaningful variable reduction in the index.

**Keywords:** Coefficient of variation, composite indices, principal component analysis

## INTRODUCTION

A composite indicator measures multi-dimensional concepts (Eg competitiveness, e-trade or environmental quality) which cannot be captured by a single indicator. Ideally, a composite indicator should be based on a theoretical framework or definition, which allows individual indicators or variables to be selected, combined and weighted in a manner which reflects the dimensions or structure of the phenomena being measured. Using composite indices we can summarize complex, multi-dimensional realities with a view of supporting decision makers. It is easier to interpret than a set of many separate indicators and can assess progress of countries over time and most importantly it is able to reduce the visible size of a set of indicators without dropping the underlying information base. In the context of policy analysis, composite indices are useful in identifying trends and drawing attention to particular issues and they can also be helpful in setting policy priorities and in benchmarking or monitoring performance. (Farrugia 2007)

On the other hand, if the composite indices are poorly constructed and the construction process is not transparent and/or lacks sound statistical or conceptual principles, it may

[1] Department of Crop Science, Faculty of Agriculture, University of Peradeniya, Sri Lanka.
[2] Postgraduate Institute of Agriculture, University of Peradeniya, Peradeniya, Sri Lanka.
* Corresponding author: priya_npd@yahoo.com

provide misleading policy messages, may invite simplistic policy conclusions and may be misused. Constructing a composite indices using a large number of variables is not practically useful. Considering the statistical relationship between variables, application of the multivariate statistical technique called principal component analysis (PCA) is a widely used technique for variable reduction. PCA involves a mathematical procedure that transforms a set of correlated variables into a smaller set of uncorrelated variables called principal components. These principal components are linear combinations of the original variables. Usually variables have to be standardized before performing PCA due to unit dependency. That means, the results of PCA depend on the scales at which the variables are measured.

Factor analysis (FA) is also a variable reduction technique and is a useful tool for investigating variable relationships for complex concepts such as socio-economic status, dietary patterns, or psychological scales. It allows researchers to investigate concepts that are not easily measured directly by collapsing a large number of variables into a few interpretable, uncorrelated underlying factors. In factor analysis, a factor is a latent (unmeasured) variable that expresses itself through its relationship with other measured variables. Variables have to be standardized to perform FA as PCA.

Since PCA operates on standardized data, scaled by their standard deviation, drawing conclusions about the dominance of variation for the actual, unstandardized data tends to be misleading. Some variables may have high inheritant variability. Another disadvantage of PC's derived using the correlation matrix is that they give coefficients for standardized variables and are therefore less easy to interpret directly (Jolliffe, 2002). Therefore this problem has to be addressed to get meaningful results from PCA.

The objective of this study is to find out the solution to the problem of unit dependency of performing PCA without standardizing the variables while preserving the information with respect to inherited variability of the variables.

## METHODOLOGY

Conducting a PCA using covariance matrix approach without standardized variables is a good alternative to preserving variability of the variables. But it has to be carried out overcoming the problem of unit dependency and influence of variables which have high variance due to relatively larger values. In order to achieve this objective the following two methods were tested.

### Method 1

Variable were converted in to new set by dividing the data of original variables by their means. In this method the new set of variables is independent of the unit and the differences between variability of each variable are substantially lower.

In this method, if the original variables are $X_i$ where *i=1, 2, …, k, where k* is the number of original variable

$Y_i = {X_i}/{\bar{X}_i}$ , where, $\bar{X}_i$ is the mean of the $i^{th}$ variable.

Then, $E(X_i) = 1$, $Var(Y_i) = (CV)^2$

## Method 2

Variables were adjusted meaningfully while keeping the variables unitless and controlling the effect of large values on higher variance. For the adjustments, some variables were expressed on a per household basis and also GN densities of population and building were divided by the total density. Then new variables were with lower variability and unitless.

In order to achieve the objectives of the study, timely important issues faced by Sri Lanka were considered. i.e. the classification of urban and rural sectors in the country. The prevailing method is not based on a proper statistical methodology. For the study, *Grama Niladhari* (GN) division level data was obtained from Department of Census and Statistics (DCS)[3]. Considering the following variables data was collected by GN divisions in Colombo District. Those variables were selected after consulting experts.

**Table 1. List of variables**

| Variable Name | Description |
|---|---|
| Sector | Urban/Rural |
| LA | Local Authority (Municipal Council/Urban Council/*Pradesheeya Sabha*) |
| GN | *Grama Niladhari* division |
| WholeS_Trade | No of wholesale trade centers |
| Retail_Trade | No. of retail trade centres (shops, groceries etc.) |
| Section_Edu | No. of education centres |
| Section_Recrn | No. of recreation centres (Cinema halls, drama theaters, Libraries, amusement parks, sports clubs, fitness centers, etc.) |
| Section_Health | Private medical centres, hospitals etc |
| Ind_above_5 | No. of industries, which the no. of employees are 5 and above |
| Pop_Density | Population density per acre and per $km^2$ |
| Mg_Emp_Pcnt | Percentage of migrated (In) population due to employment out of the population in GN |
| Mg_Edu_Pcnt | Percentage of migrated (In) population due to education out of the population in GN |
| HH_Density | No. of households per acre and per $km^2$ |
| Stories_more2_pcnt | Percentage of more than two storied houses with respect to total no. of houses in the GN (as a proxy variable for land value) |
| Hh_Bildn_Pctn | Percentage of houses with respect to total no. of buildings in the GN |
| Buildn_Density | No. of buildings per acre and per $km^2$ |

Data was obtained from two censuses conducted by the Department of Census and Statistics. Population, migration, Housing and building related data was taken from the Population and Housing census conducted in 2012 and establishment related data was taken from the listing stage of Economic Census conducted in 2013. In Colombo district, 557 number of GNs were considered for the study.

[3] All the centres belong to the government are not included

## RESULTS AND DISCUSSION

Nature of variables was identified using descriptive statistics. Set of variables with remarkably different variability can be seen from the table 1. That was not only due to magnitude of the numbers but also due to inherent property of the variable. As an example the variable Mg_Emp_Pcnt (Percentage of migrated people due to employment) varied from 0.2 to 93.7 with the variance of 46.8. But the variable, HH_Bildn_Pctn (Percentage households) ranged from 0.2 to 97.3 while showing more variance (124.4).

**Table 2. Descriptive statistics of the considered variables**

| Variable | Minimum | Maximum | Mean | Variance | CV |
|---|---|---|---|---|---|
| WholeS_Trade | 0.00 | 184.00 | 8.35 | 288.84 | 2.04 |
| Retail_Trade | 0.00 | 597.00 | 69.40 | 5700.34 | 1.09 |
| Section_Edu | 0.00 | 63.00 | 9.61 | 83.12 | 0.95 |
| Section_Recrn | 0.00 | 33.00 | 3.42 | 21.60 | 1.36 |
| Section_Health | 0.00 | 35.00 | 4.47 | 23.15 | 1.08 |
| Ind_above_5 | 0.00 | 43.00 | 7.58 | 48.40 | 0.92 |
| Pop_Density (per acre) | 0.31 | 204.70 | 25.83 | 969.73 | 1.21 |
| Mg_Emp_Pcnt | 0.22 | 43.03 | 10.09 | 34.28 | 0.58 |
| Mg_Edu_Pcnt | 0.00 | 31.58 | 2.21 | 9.74 | 1.41 |
| HH_Density per acre) | 0.08 | 45.77 | 6.12 | 45.52 | 1.10 |
| Hh_Bildn_Pctn | 6.17 | 97.26 | 76.15 | 114.27 | 0.14 |
| Buildn_Density (per acre) | 0.10 | 74.57 | 8.33 | 90.01 | 1.14 |
| Stories_more2_pcnt | 0.00 | 10.47 | 1.58 | 3.41 | 1.17 |

As a variable reduction technique, PCA was performed to identify the minimum linear combination of considered variables with higher explanation of the original variation of the data. Initially this was conducted using correlation matrix approach to overcome unit dependency problem. Number of GNs was 557.

### Application of PCA - Correlation matrix approach

Considering the results of PCA, first four PCs with eigen values are above 1, explained only 80 percent of total variation. This is not supposed to be a good approach due to two reasons. One of these was requirement of selecting higher number of PCs though the objective is to reduce variables to a minimum level while explaining the greater degree of variability whereas the other was neglecting the inherent variability due to standardizing variables. Then the possible alternative is to perform PCA using covariance matrix approach.

### Covariance matrix approach

The main disadvantage of covariance matrix approach is the scale dependency. In order to test the effect of unit on PCs, separate PCA was performed taking the variables relating to densities in two separate units, one in per acre and the other per square kilometer. With unit as per acre, first PC explained 83.47 percent of total variation, while using densities as per $Km^2$ 98.7 percent of total variation was explained by the first PC. The clear difference in percentage is an indication that depending on unit PCA outcome is affected. In addition, with both units large values for the percentage indicate covariance approach instead of correlation approach (after standardizing the variables) should be used in constructing PCs. Moreover in

the factor analysis, nine variables have significantly contributed to the first factor with densities unit per acre whereas only 3 variables were significant under the unit as per $Km^2$. This also indicates, variables in different units provide different results. Therefore unit dependency has to be removed without loss of variability. As solutions to the problem given by two methods were tested.

**Method 1: Coefficient of Variance (CV) method**

To meet the objective variables were transformed by dividing by their means. Then unit dependency problem was solved and the inherent variability of the variables were also taken into account. Then the variances of the new set of variables are the square term of CV of original variables. Table 2 shows that difference between the CV values which are comparatively lower than variances of the original variables.

**Table 3. Eigen values of both Correlation and Covariance matrix approaches**

| PC No. | Correlation matrix approach | | Covariance matrix approach | |
|---|---|---|---|---|
| | Eigen value | Cum. Pcnt. of variance(%) | Eigen value | Cum. Pcnt. of variance(%) |
| 1 | 5.44 | 41.82 | 8.20 | 46.13 |
| 2 | 2.52 | 61.20 | 2.84 | 62.11 |
| 3 | 1.30 | 71.24 | 2.40 | 75.64 |
| 4 | 0.78 | 77.26 | 1.69 | 85.17 |
| 5 | 0.72 | 82.76 | 0.79 | 89.62 |
| 6 | 0.55 | 86.96 | 0.56 | 92.79 |
| 7 | 0.52 | 90.93 | 0.49 | 95.55 |
| 8 | 0.43 | 94.28 | 0.30 | 97.22 |
| 9 | 0.30 | 96.60 | 0.24 | 98.55 |
| 10 | 0.22 | 98.30 | 0.19 | 99.64 |
| 11 | 0.19 | 99.73 | 0.04 | 99.86 |
| 12 | 0.02 | 99.91 | 0.02 | 99.96 |
| 13 | 0.01 | 100.00 | 0.01 | 100.00 |

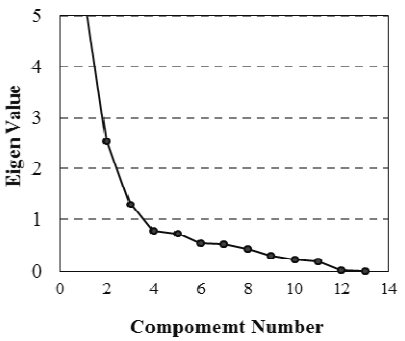


**Fig. 1. Scree plot for Correlation matrix approach**

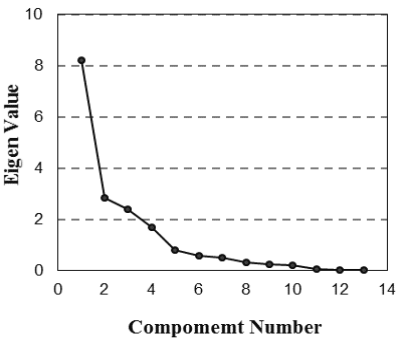


**Fig. 2. Scree plot for Covariance matrix approach**

In correlation approach 71.24 percent of total variance was explained by the first three factors, while 75.64 percent was explained with the covariance approach. This is about 4.5 percent of improvement which cannot be considered as sufficient. Scree plot also do not show a clear

difference. In factor analysis, without rotation, in covariance matrix approach, one variable indicated significant contribution on two factors. Therefore Varimax rotation was used to overcome the problem. In correlation matrix approach 12 variables out of 13, significantly contributed to first three factors while in covariance matrix approach it was 11. Since the application of CV method, a significant improvement on covariance approach over the correlation approach could not be seen, the second method was tested.

**Method 2: Variable adjustment method**

The variables, WholeS_Trade, Retail_Trade, Section_Edu, Section_Recrn and Section_Health were adjusted dividing the values by total number of households in the GN division. The adjusted variables were named as WS_HH, Retail_HH, Edu_HH, Recrn_HH, and Health_HH. The rational of this adjustment is that in urban area more outside people of the GN division get benefited from wholesale and retail trade centres, education, and health and recreation centres.

When the value of adjusted variable in a GN division is high, that area is more likely to be urban. Five variables were adjusted to the total number of households. It is not necessary to consider another variable for household density since that information is reflected in the adjusted variables. Therefore two variables of population density and building density were adjusted to those of total densities for the whole country. Then the variables became unitless and they were renamed as pop_tot_density and bld_tot_density. The rational of this adjustment is that when the density of the considered variable in a GN division with respect to whole country density is high, the particular area has more urban nature. Before adjustment, there are very huge differences in variances. But after adjustment, variables became unitless and the differences of variances among variables are comparatively very low. Hence, using adjusted variables, PCA was performed.

**Table 4. Eigen values of both Correlation and Covariance matrix approaches**

| PC No. | Correlation matrix approach | | Covariance matrix approach | |
|---|---|---|---|---|
| | **Eigen value** | **Cum. Pcnt. of variance (%)** | **Eigen value** | **Cum. Pcnt. of variance (%)** |
| 1 | 3.97 | 33.12 | 1270.30 | 87.58 |
| 2 | 2.42 | 53.31 | 126.94 | 96.34 |
| 3 | 1.57 | 66.38 | 28.18 | 98.28 |
| 4 | 1.04 | 75.02 | 14.85 | 99.30 |
| 5 | 0.94 | 82.85 | 7.73 | 99.84 |
| 6 | 0.55 | 87.43 | 1.89 | 99.97 |
| 7 | 0.45 | 91.19 | 0.49 | 100.00 |
| 8 | 0.39 | 94.42 | 0.00 | 100.00 |
| 9 | 0.34 | 97.21 | 0.00 | 100.00 |
| 10 | 0.17 | 98.63 | 0.00 | 100.00 |
| 11 | 0.14 | 99.81 | 0.00 | 100.00 |
| 12 | 0.02 | 100.00 | 0.00 | 100.00 |

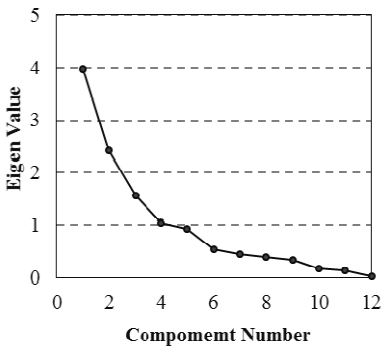


**Fig. 3. Scree plot for Correlation matrix approach**

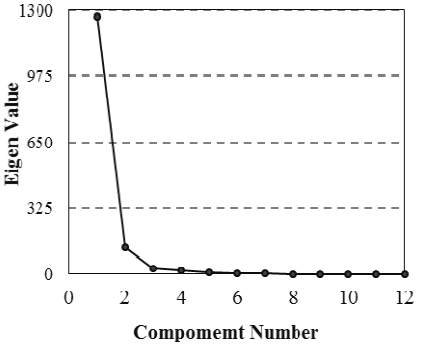


**Fig. 4. Scree plot for Covariance matrix approach**

According to the table 4, 96.34 percent of total variance was explained by the first two factors in covariance matrix approach. Scree plot also clearly shows that. But to reach that proportion, in correlation matrix approach 8 factors was needed. That is unacceptable. Therefore, the covariance matrix approach is more useful. Under the factor analysis without rotation, in correlation matrix approach, one variable indicated significant contribution on two factors and it was solved using Equifax rotation. According to the results, in correlation matrix approach 11 variables out of 12, significantly contributed to the first four factors while in covariance matrix approach it was 7 for the first two factors.

## CONCLUSIONS

Constructing a composite index using a minimum number of meaningful variables is very useful. Principal component analysis (PCA) is a widely used technique for variable reduction. Conducting PCA as a variable reduction techniques, with correlation matrix approach is not always acceptable due to ignorance of the inherent variability of variables. Covariance matrix approach is a good solution to the above problem, but it has the drawback of unit dependency. To overcome that issue out of the two methods studied, the second method can be used. With this method, adjusting variables meaningfully while keeping them unitless and controlling the effect of large values on higher variance gives solution for the issues mentioned above. Therefore adjusting variables in constructing composite indices can generate very useful information.

## REFERENCES

Silva, A.P.G.S. (2000) Construction of Human Development Indices using Multivariate Techniques, PGIA, University of Peradeniya

Jolliffe, T. (2002) Principal Component Analysis, Springer Verlag, New York

Josseph, F., William C.B., Barry J.B. and Rolph E.A. (2010) Multivariate Data Analysis. A Global Perspective, Pearson Education Inc, New Jersey

Fernando, S., Samita, S. and Abenayake, R. (2011). Modified factor analysis to construct composite indices: Illustration on Urbanization index. Trop. Agric. Res., *24*, 271 - 281

Tuan, A. and Magi, S. (2009) Principal Component Analysis: Final Paper in Financial Pricing. National Cheng Kung University, 3-26

Farrugia. N, (2007) conceptual issues in constructing composite indices. Occasional papers on islands and small states, 2/2007, 2-36

Delchambre, L. (2014) Weighted principal component analysis: a weighted covariance eigen decomposition approach, University of Liege, Belgium, 1-5