American Journal of Applied Mathematics and Statistics, 2021, Vol. 9, No. 3, 75-82
Available online at http://pubs.sciepub.com/ajams/9/3/1
Published by Science and Education Publishing
DOI:10.12691/ajams-9-3-1

Modelling a Multilevel Data Structure Using a Composite Index

Prabath Badullahewage*, Dilhari Attygalle

Department of Statistics, University of Colombo, Colombo, Sri Lanka
*Corresponding author: prabathbadullahewage@gmail.com

Received June 06, 2021; Revised July 09, 2021; Accepted July 23, 2021

Abstract When modelling complexed data structures related to a certain social aspect, there could be various hierarchical levels where data units are nested within each other. There could also be several variables in each level, and those variables may not be unique for each case or record, making the data structure even more complexed. Multilevel modelling has been used for decades, to handle such data structures, but may not be effective at all times to capture the structure fully, due to the extent of complexities of the data structure and the inherent issues of the procedure. On the contrary, ignoring the multilevel data structure when modelling, can lead to incorrect estimations and thereby may not achieve acceptable accuracies from the model. This research explains a simple approach where a complexed multilevel structure is compressed to a single level by combining higher level variables to form a composite index. Moreover, this composite index, also reduces the number of variables considered in the entire modelling process, substantially. The process is exemplified, using a primary data set gathered on household education expenditure using a systematic sampling survey. Several variables are collected on each household and another set of variables relating to each school going child in the household, creating a multilevel data structure. The composite index, named as, "Household Level Education Index" is developed through a factor analysis where the detailed process of its construction is explained. The LASSO regression was performed to illustrate the use of the proposed composite index by predicting the monthly household education expenditure through a single level regression model. Finally, a Random Forest model was used to examine the feature importance, where the proposed composite index "Household level education index" was the most important feature in predicting the monthly household educational expenditure.

Keywords: *composite index, multilevel modeling, factor analysis, educational expenditure*

Cite This Article: Prabath Badullahewage, and Dilhari Attygalle, "Modelling a Multilevel Data Structure Using a Composite Index." *American Journal of Applied Mathematics and Statistics*, vol. 9, no. 3 (2021): 75-82. doi: 10.12691/ajams-9-3-1.

1. Introduction

Data and Information is a hot topic in the modern era as it is the key factor which makes the strategic decision making a success. In the recent decades the world has witnessed a massive escalation in available information and the founder of World Economic Forum describes the extent and the use of this upsurge of data as the "Fourth Industrial Revolution" [1]. Availability of information in large scale comes with a price and thus needs to be interpreted and consolidated effectively and meaningfully. The general public or any entity who is interested in such data face difficulties when they are presented with a wide range of indicators to encompass all the necessary information on a particular phenomenon. Instead, if the available information is presented in the form of a sole number that incorporates this wide range of indicators, then it will be much easier for the audience to grasp a complex concept. Formulating a Composite Index is a technique that has been used to reflect many social phenomena and is usually a sole number.

Lately, Composite Indicators have acquired incredible popularity in a wide range of research areas. The use of composite indicators by a vast number of worldwide organizations has caught the attention of the media and policymakers around the world, and their utilizations have increased from that point forward [2]. According to [3], Composite Indicators have gained immense popularity in a variety of fields and Reference [3] has pointed out over 400 official composite indicators that rank a nation based on financial, political, social, or environmental measures. Moreover, [4] has reported on more than 100 composite measures of human progress in a complementary report by the United Nations' Development Programme.

The idea of using a composite index has an interdisciplinary nature, hence it can be applied into many research areas [5]. However, the use of a Composite index when modelling multilevel data structures are rarely spoken. The multilevel data structures are encountered when data are being collected on various needs, for

example on a social phenomenon or in a biological setup. For instance, children are nested within families and families are nested within neighbourhoods whereas in biological studies, patients are nested within hospital. It is evident that such multilevel data structures are commonly available in many data structures collected in various other phenomena, and thus need adequate attention when analysing or modelling such data.

One of the most common modelling approaches when dealing with multilevel data structures is, multilevel modelling. However, multilevel modelling has its own drawbacks, and could make the modelling process complex. Multilevel models are data and theory intensive and also based on many assumptions about data [6]. These assumptions can be demanding at times, and if not met, then it will affect the statistical inference from such models, adversely. On the other hand, in most of the modelling procedures, variable selection is essential in building a good simple model. The variable selection procedure in a multilevel data structure, however, must be done with caution, as predictors can be selected for each level of the model and interactions between predictors can be considered at either level or across levels [7]. The same article, written by reviewing 98 selected journals between 1999 and 2003, describes that there are some underlying issues in the reporting of multilevel data analysis.

This research is focused on proposing an alternative method that can be used to bring the multilevel data to a single level using a composite index. The procedure is explained using an example data set, gathered at two hierarchical levels, where the information at the higher level is combined in creating a composite index reflecting those variables in that level. This composite index will bring all the information required for modelling into a single level, while also reducing the number of predictors substantially. The use of this approach will thus simplify the complexity in a multilevel data structure, while serving as an alternative approach, to a typical multilevel modelling procedure.

In this study a data set regarding the household education expenditure was used and the data are in two levels namely household level and the child level. The study will illustrate how to bring the child level information regarding education expenditure, to the household level, using the proposed composite index methodology. Formulating this index will result in all variables being in one level where a suitable model can then be applied. The education expenditure of the households is modelled using the proposed composite index as a single predictor, along with the other predictors, as a solution to the issues related to modelling multilevel data structures.

2. Methodology

2.1. Creating a Composite Index

In this study a Composite Index was used to bring the multilevel data structure to a single level. It is worth noting that, due to the nature of the data in this example, all variables in one level were associated with children's education, and hence were used to create the composite index. Creating a composite index is a method of summarizing a set of variables by combining the individual variables [8]. The study used this method to summarize seven variables regarding the school going children in the household. Number of variables per household regarding the children was not unique as the particular household may have varying number of school-going children. It should be noted that the primary focal point of the study was the household, and not the individual child.

Prior to the composite index building, it is essential to normalize the individual variables involved in this process [8]. They also stated that "Normalization is required prior to any data aggregation as the indicators in a data set often have different measurement units. Therefore, it is necessary to bring the indicators to the same standard, by transforming them into pure, dimensionless, numbers." Min-Max normalization was used in this study and R software was used to do the transformation. The Min-Max normalization equation is defined as follows [9].

$$X_{\min-\max} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

where "x" is the individual variable under composite index creation.

There are different methods to construct a composite index and it should be noted that there is no universally best method to be used at all times. In each case the construction of an index is much determined by the particular application, by incorporating some expert knowledge on the phenomenon and the problem of interest [8].

This study used a Factor Analysis (FA) based method to construct the composite index [9]. Factor analysis groups together individual variables which are collinear to form a composite indicator. The composite indicator captures as much as possible of the information common to individual variables [10]. At first, FA was performed in order to get the Factor scores. Eigen values greater than one rule was used to find the number of factors to be retained and the Varimax rotation was used to load the individual variables to the selected factors. The resulting factor scores were computed and were treated as intermediate composite indicators.

Consequently, these intermediate composites were aggregated by assigning a weight to each of them equal to the proportion of the explained variance in the data set. The aggregated intermediate composite indicator was taken as the final composite index. The following equation (2) was used in constructing the composite index [9,11].

$$Composite\ Index = \frac{\sum_{i=1}^{k} f_i v_i}{\sum_{i=1}^{k} v_i} \quad (2)$$

$$v_i = \frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i} \quad (3)$$

where;
k = the number of factors selected
f_i = factor score of the i^{th} factor
v_i = the proportion of variance explained by i^{th} factor (the weight of the intermediate composite indicator)

λ_i = eigen value of the i[th] factor

The created composite index captures the information in a multilevel structure and brings them to a single level as a numeric variable. The composite index was used as a continuous explanatory variable in the statistical modelling.

2.2. Statistical Modelling

LASSO regression was used to model the data as it is a method to overcome the issue of multicollinearity without omitting the predictor variables. It also performs the variable selection. LASSO is a Shrinkage method as it shrinks the coefficient estimates towards zero. The LASSO coefficients (β_j) are estimated by minimizing the following quantity [12] given in equation (4),

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \quad (4)$$

where ;

y_i = actual value of the i[th] observation

P = number of parameters

λ = tuning parameter, $\lambda \geq 0$.

The term $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$ is called LASSO penalty. If the tuning parameter λ is sufficiently large, then LASSO will set some of the coefficients exactly to zero which leads to a variable selection. Best value of λ can be found by K-fold cross validation.

Random forest regression was also used in the study to complement the modelling approach especially with regard to elaborate on the feature importance of the model. It is a method that gives the predictions by constructing multiple regression trees. It usually builds trees on bootstrapped training samples and does the prediction on all the trees. When building these regression trees, each time a split in a tree is considered, a random sample of predictors is chosen as split candidates. The final prediction is taken as the average of all the regression tree predictions. Random forests can be used to rank the importance of variables in a regression model in a natural way. This variable importance was used in this study to showcase the success of the proposed composite index in the LASSO regression model [12].

3. Analysis

3.1. Creating the Composite Index

This study uses a primary data set containing household education expenditure and several other variables that are potential determinants of it. The data were obtained from a carefully designed survey which was carried out in one Grama Niladhari (GN) division in the Homagama City, in Sri Lanka. Data from 239 households were collected using a systematic sampling technique. Face to face interviews on a pre-prepared questionnaire was used as the data collection method. The fact that most of the educational research involved with the hierarchical data structure and multilevel modeling approaches [13] motivated this study to use a primary data set containing household education expenditure.

This study has gathered data on both household level and child level. Information regarding each child in a house is captured under child level data. This child level data is anticipated to be associated with household education expenditure. Child level data is consisted of the following seven variables that vary with every child in a household.

- Grade
- School type
- Absenteeism
- Performance
- Parents' level of satisfaction regarding child's performance
- Parents' level of satisfaction regarding child's school
- Education goal

Absenteeism and performance variables are continuous variables while the other five variables were captured on an ordinal scale. This child level information is not unique due to the fact that each house has varying number of school-going children. One of the most difficult challenges faced by the educational statistics researchers is to incorporate micro and macro information into one statistical model [14]. That is, in order to carry out the analysis, it was required to combine the individual level information with the information of the groups they belonged to. In this research too, it was required to combine child level data with the household level data.

With this example data set, a method was needed to bring the varying child level information to the household level, to reflect the educational status of school going children of a household. It must be noted that the individual data coming from each child are not possible to be incorporated into a model due to the structural variation those variables possess. As a suitable remedy, a Composite Index was created to overcome this issue. The Composite index will bring the child level data into household level and by doing so the two-level hierarchical structure of the data set can be reduced down to one level. Hence, the proposed approach can overcome the said issues with regard to the data structure and also bypass the multilevel modelling approach.

The study noted that the maximum number of children observed per household was three. It was found that there was no significant difference in the mean of monthly household expenditure on education between those houses with two and three children. Also, the number of houses having three children is 15 and it is comparatively low. T-test was used to do the above mean comparison and the test was not significant at 5% significance level. Hence, houses with two and three children were taken together when creating the composite index. Accordingly, the composite index was created separately for houses with one child and houses with two or three children.

Min-Max normalization was applied on variables before creating the composite index to bring all the variables to the same standards and was done using R software.

According to the data set there were 239 households in total. Out of the 239 households, 130 households had one child while rest of the 109 households had two or three children. Factor Analysis was first applied to houses with one child and two factors had an Eigen value greater than

one. Those factors were extracted, and results are shown in the Table 1 and Table 2.

Table 1 shows the total variance explained by the Principal Component Analysis (PCA) for houses with one child while Table 2 shows the Varimax Rotated component matrix for houses with one child.

Table 1. Total Variance Explained by the PCA for Houses with one Child

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.424	34.631	34.631
2	1.549	22.127	56.758
3	0.959	13.694	70.453
4	0.76	10.856	81.309
5	0.546	7.793	89.102
6	0.42	6.003	95.104
7	0.343	4.896	100

Table 2. Varimax Rotated Component Matrix for Houses with one Child

Variable	Component	
	1	2
A	0.025	0.843
B	0.458	0.715
C	-0.379	0.555
D	0.741	-0.171
E	0.624	0.032
F	0.81	0.049
G	0.67	0.117

Note that for the simplicity of representation, the seven variables in the Table 2 are coded from A-G. Refer Table 3 to check the variable name corresponding to the code.

Table 3. Variables Names and their Codes

Code	Variable name
A	Child 01 Grade
B	Child 01 School
C	Child 01 Absenteeism
D	Child 01 Performance
E	Child 01 Parents School Satisfaction
F	Child 01 Parents Performance Satisfaction
G	Child 01 Education Goal
H	Child 02 Grade
I	Child 02 School
J	Child 02 Absenteeism
K	Child 02 Performance
L	Child 02 Parents School Satisfaction
M	Child 02 Parents Performance Satisfaction
N	Child 02 Education Goal

The factor scores were computed as follows and are given in the equations (5) and (6). These factor scores were treated as the intermediate composites.

$$\begin{aligned} F1 = {} & 0.741 \times Child\ 01\ performance \\ & + 0.624 \times Child\ 01\ school\ satisfaction \\ & + 0.810 \times Child\ 01\ performance\ satisfaction \\ & + 0.670 \times Child\ 01\ education\ goal \end{aligned} \quad (5)$$

$$\begin{aligned} F2 = {} & 0.843 \times Child\ 01\ grade \\ & + 0.715 \times Child\ 01\ school\ type \\ & + 0.555 \times Child\ 01\ absenteeism \end{aligned} \quad (6)$$

The corresponding Eigen values were 2.424 and 1.549, respectively. The composite index for houses with one child was calculated as follows using the equation (2) explained under methodology.

$$Composite\ Index = \frac{2.424}{(2.424+1.549)} \times F1 + \frac{1.549}{(2.424+1.549)} \times F2 \quad (7)$$

Similarly, the method was applied to the houses with two or three children and the results are shown in the Table 4 and Table 5.

Table 4. Total Variance Explained by the PCA for Houses with two or three Children

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.414	24.387	24.387
2	1.896	13.54	37.928
3	1.507	10.766	48.694
4	1.418	10.128	58.822
5	1.285	9.179	68
6	1.069	7.635	75.635
7	0.738	5.271	80.906
8	0.671	4.795	85.701
9	0.545	3.892	89.593
10	0.45	3.215	92.808
11	0.367	2.62	95.428
12	0.271	1.939	97.367
13	0.219	1.564	98.931
14	0.15	1.069	100

Table 5. Varimax Rotated Component Matrix for Houses with two or three Children

Variable	Component					
	1	2	3	4	5	6
A	-0.153	0.12	0.081	0.095	0.159	0.773
B	0.139	0.135	0.869	0.064	0.165	0.013
C	-0.032	0.045	-0.005	0.849	0.067	0.119
D	0.71	0.064	0.406	-0.093	-0.075	-0.261
E	0.214	0.819	0.057	-0.076	-0.158	-0.029
F	0.495	0.472	0.279	-0.388	-0.231	0.036
G	0.878	0.017	0.085	0.056	0.134	0.059
H	0.264	-0.196	0.12	-0.119	-0.265	0.759
I	0.134	-0.043	0.84	-0.092	0.136	0.203
J	0.089	-0.123	-0.004	0.785	-0.257	-0.106
K	0.196	-0.076	0.258	-0.088	0.775	-0.117
L	-0.11	0.841	-0.014	0.039	0.241	-0.011
M	0.003	0.409	0.19	-0.366	0.552	0.167
N	0.699	0.075	-0.011	0.149	0.503	0.206

Note that for the simplicity of representation, the 14 variables in the Table 5 are coded from A-N. Refer Table 3 to check the variable name corresponding to the code. The corresponding factor scores (intermediate composites) were computed as follows,

$$F1 = 0.710 \times Child\ 01\ performance + 0.495 \times Child\ 01\ performance\ satisfaction + 0.875 \times Child\ 01\ education\ goal \quad (8)$$

$$F2 = 0.819 \times Child\ 01\ school\ satisfaction + 0.841 \times Child\ 02\ school\ satisfaction \quad (9)$$

$$F3 = 0.869 \times Child\ 01\ school\ type + 0.840 \times Child\ 02\ school\ type \quad (10)$$

$$F4 = 0.849 \times Child\ 01\ absenteeism + 0.785 \times Child\ 02\ absenteeism \quad (11)$$

$$F5 = 0.775 \times Child\ 02\ performance + 0.552 \times Child\ 02\ performance\ satisfaction + 0.503 \times Child\ 02\ education\ goal \quad (12)$$

$$F6 = 0.773 \times Child\ 01\ grade + 0.759 \times Child\ 02\ grade \quad (13)$$

The respective eigen values were 3.414, 1.896, 1.507, 1.418, 1.285, and 1.069. Composite index for houses with two or three children were calculated as follows,

$$\begin{aligned} Composite\ Index = {} & \frac{3.414}{(3.414+1.896+1.507+1.418+1.285+1.069)} \times F1 \\ & + \frac{1.896}{(3.414+1.896+1.507+1.418+1.285+1.069)} \times F2 \\ & + \frac{1.507}{(3.414+1.896+1.507+1.418+1.285+1.069)} \times F3 \\ & + \frac{1.418}{(3.414+1.896+1.507+1.418+1.285+1.069)} \times F4 \\ & + \frac{1.285}{(3.414+1.896+1.507+1.418+1.285+1.069)} \times F5 \\ & + \frac{1.069}{(3.414+1.896+1.507+1.418+1.285+1.069)} \times F6 \end{aligned} \quad (14)$$

According to Table 1, the first two components together explain around 56.76% of the variation in the data for the household with one child. The two-factor model was selected for the households with one child by adhering to the eigen values greater than one rule. But if this proposed methodology is to be applied in a practical situation, then one can even go for a three-factor model as the third component is nearly equal to one and by doing so three factor model will explain around 70.45% of the variation in the data. According to Table 3, the first six components together explain approximately around 75.64% of the variation in the data for the households with two or three children. The researchers anticipate that this amount is sufficient for further analysis of the data. Hence, the fitted factor models are adequate.

The composite indices calculated by the equations (7) and (14) were considered as a single numerical predictor variable when modelling educational expenditure. The final composite index variable is named as "Household Level Education Index". The varying child level information in a particular household is captured under the "Household Level Education Index". Moreover, the multilevel structure of the data is reduced to a single level and a considerable data reduction is also achieved. The index scores vary from 0.42 to 2.16 of which the distribution is shown in Figure 1.

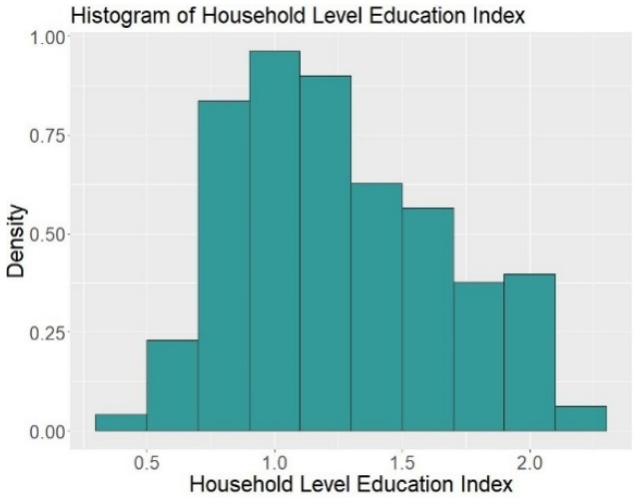

Figure 1. The distribution of Household Level Education Index

3.2. Statistical Modeling

Recall that the Household Level Education Index has brought the child level information, regarding education expenditure, to the household level while reducing the number of variables in the study. The monthly household education expenditure was modeled using Household Level Education Index as a continuous explanatory variable along with the other variables. The LASSO regression modelling was used in this study as a remedy for the multicollinearity issue. The distribution of the response variable (monthly household educational expenditure) was positively skewed and thus far away from normality. As a solution, log (natural logarithm) transformation was applied on the response variable to get rid of the non-normality. Log transformed response variable with all the explanatory variables were used to fit the LASSO model. The LASSO model was fitted using R software.

The dataset was divided randomly into training data and testing data prior to model fitting. Training data set consisted of 80% data while testing data sets consisted of 20% data. Training data set was used in model fitting and parameter estimation, while testing data set was used to evaluate prediction accuracy. Next step was to select the best lambda value for the LASSO model. A 10-fold cross validation procedure was used to obtain the Mean Squared Error (MSE) for each respective lambda value with the training dataset in obtaining the best lambda value for LASSO regression model.

Best lambda (which gives the minimum cross validated MSE) was 0.001293 and it had 29 variables (including dummy variables). The lambda under one standard deviation was 0.033557 and it had 13 variables (including dummy variables). The best lambda gave the lowest test MSE. Hence, best lambda (0.001293) was selected as the optimal lambda to get higher prediction accuracy. A LASSO model with lambda set to 0.001293 was fitted as the final LASSO model.

The final LASSO model is given in the Table 6 as predictors and the estimated regression coefficients. The dependent variable of the final LASSO model is monthly household education expenditure.

Table 6. The Final LASSO Model

Predictor	Estimated Regression Coefficient
Number of Household members	0.004
Number of male children	0.559
Number of female children	0.606
Religion_Christian	0.816
Household head gender_Female	-0.345
Household head job category_2	-0.008
Household head job category_3	0.034
Household head job category_4	-0.033
Household head job category_5	0.235
Spouse job category_2	-0.159
Spouse job category_3	-0.043
Spouse job category_5	0.083
Spouse job category_6	-0.451
Household head education level_4	0.222
Spouse education level_2	-0.014
Spouse education level_4	-0.064
parent status of household_both	-0.254
Monthly income_3	0.071
Monthly income_4	0.418
Monthly income_5	0.434
Other income_yes	0.080
Household debt_2	-0.217
Household debt_4	-0.228
Household debt_5	-0.014
Household debt_6	0.203
Scholarship_yes	0.210
Household Level Education Index	0.548
constant	-0.014

It should be noted that the predicted monthly household education expenditure is in log format. Exponentials of the predictions must be taken to get them in original format. The prediction accuracy of the test set was measured using Root Mean Squared Error (RMSE). The test set accuracy measure for the final LASSO model in original format is as follows,

- RMSE = 11 580.69

The dependent variable being measured in Rupees and having large expenditure amounts (exceeding Rs 10 000 most of the time) is the reason for having large values for accuracy measures in original format. It should be noted the that the prediction accuracy improvement is not an objective of this study but is mainly focused on explaining a simple approach where multilevel data structure is compressed into a single level using a composite index.

According to the LASSO regression model given in Table 6, The proposed composite index "Household Level Education Index" is a significant variable with a moderately large positive regression coefficient. The variable "Household Level Education Index" is incorporated in the model as a single continuous variable. The information about the higher level (child level information) is incorporated into this "Household Level Education Index" and it is being used at the lower level (household level) as a separate variable. Multilevel modeling was not required in this approach as the hierarchical data structure was simplified into a single structure using the "Household Level Education Index". Hence, it is evident that the proposed composite index methodology is an alternative and a simple method for modeling multilevel data structures.

Random forest regression was applied on the original data to obtain the feature importance in predicting the monthly household expenditure on school education. It was interesting to note that the significant variables in the LASSO model aligns with the features extracted as important in the random forest model.

Random forest model was fitted using 200 trees and seven variables at each split. The above two hyper parameter values were used as they gave the optimal results. The resulting test set RMSE was 13 271.91. The Figure 2 shows the feature importance plot for the above random forest model.

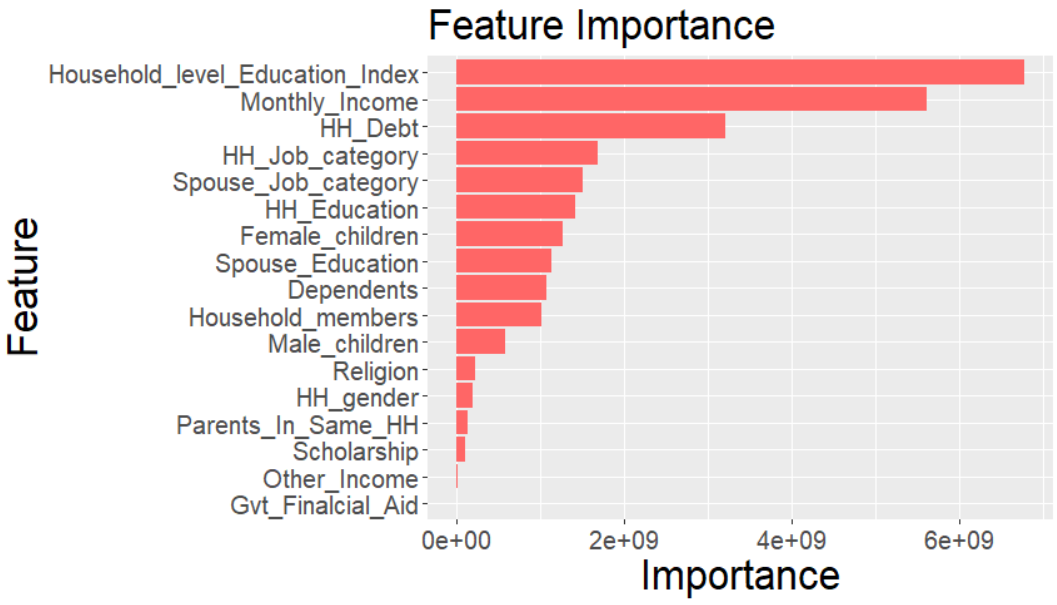

Figure 2. Feature importance of the Random Forest model

According to the feature importance given in Figure 2, the proposed composite index which is named as the "Household Level Education Index" has the highest importance in predicting the monthly household education expenditure. The proposed composite index being in the top of the feature importance is an indication of the successfulness of combining all variables in one level through the proposed composite index.

The use of composite index called "Household Level Education Index" paved the way to bring the multilevel structure of the education data into single level and model the education expenditure of households as one model using LASSO regression. The feature importance of the Random Forest model complemented the proposed methodology as the highest importance was achieved by the "Household Level Education Index" when predicting the monthly household education expenditure.

4. Discussion

This paper is aimed at presenting an alternative approach when modelling a multilevel data structure, using a composite index. The composite index was used to bring the multilevel data structure into a single level which is the main level of interest, so that a usual modeling procedure can be applied easily. As the multilevel models are data and theory extensive while heavily depending on the assumptions [6], this paper presents an alternative method, capable of bypassing these limitations. In the example used in this paper, the multilevel data structure was modelled using a LASSO regression model coupled with a composite index. A second advantage of the proposed approach is the reduction of explanatory variables used in modelling. In addition, with our example dataset, the variables related to every household was not unique, and therefore will encounter issues in modelling. This problem too was dealt with, by using a composite index successfully.

In general, the composite index will reduce the number of variables to be considered in modelling substantially, and therefore address model flexibility and curtail the complexities encountered in a multilevel data structure.

This study used a data set which contains information regarding monthly household educational expenditures and its possible determinants. It is worth noting that there was a maximum of three school going children per household in the sample. The composite index was created separately for the houses with one child and for the houses with two or three children assuming there are only maximum of three school going children per household in the whole GN division. Creating the composite index this way eliminated the problem of having varied number of variables pertaining to each households.

The hierarchical nature of the aforesaid data set was suitable for illustrating the method, as it consisted of two levels called child level and household level, having a complex data structure. However, this proposed approach is not limited only to data structure with two levels such as the household expenditure data set but can be applied suitably to other hieratical data structures, giving promising results.

The composite index, although widely used to explain various phenomena, is rarely used nor highlighted when modelling multilevel data structures. This research has exemplified the value of a composite index when modelling a multilevel data structure. When applied to the educational expenditure dataset, a satisfactory result was shown with a high accuracy of the model where the proposed composite index appearing to be the most important feature in predicting educational expenditure.

5. Conclusion

The results showed that the application of the proposed method on the monthly household educational expenditure was satisfactory as the proposed Composite Index turned out to be the most important feature in predicting expenditure while reducing the multilevel data structure into a single level. In addition, the method helped to reduce the number of variables in the study and also accommodated the problem of having a non-unique set of variables pertaining to each household. Hence, the Composite Index based method can be considered as an alternative method when modelling certain complexed multilevel data structures. The proposed method will bring the higher level information into the lower level and it will allow the researcher to model the data at the lower level without considering the hierarchical structure. This will allow the researcher to bypass the use of multilevel modeling while avoiding the disadvantages and limitations the approach possess. The proposed method can be suitably extended to hierarchical data structures with more than two levels as well. Among many other complex methods to create a composite index to summarize the information, an alternative method is to use Multi Criteria Analysis (MCA). This research however has proposed a useful concept and an alternative approach, when modelling complexed multilevel data structures.

Acknowledgements

We acknowledge Ms.Buddhi Tharanga Karunasena (Divisional Secretary-Homagama Divisional Secretariat, Sri Lanka) and Mr. Sanjaya Somarathna(Sub Inspector of Police at Homagama Police Station, Sri Lanka) who helped us with granting the approvals to carry out the survey in Homagama area. Special thanks goes out to Mrs.T.A.Chamini Gunarathne (Grama Niladhari of Homagama West, Sri Lanka), Mr. Kasun Madhushanka, Mr. Denuka Dissanayake, Mr. Lahiru Siriwardena, Mr. Janaka Attanayake and Mr. Thilanka Perera who helped us with the data collection process and for the immense support and guidance provided throughout the research.

Abbreviations

FA : Factor Analysis
GN : Grama Niladhari
PCA : Principal Component Analysis
RMSE : Root Mean Squared Error
MCA : Multi Criteria Analysis

References

[1] Schwab, K. (2016). The fourth industrial revolution: What it means and how to respond. Retrieved from https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-howto-respond. Accessed 25 March 2021.

[2] Greco, S., Ishizaka, A., Tasiou, M. *et al.* On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. *Soc Indic Res.* 141, 61- 94 (2019).

[3] Bandura, R. (2011). Composite indicators and rankings: Inventory 2011. Technical report, Office of Development Studies, United Nations Development Programme (UNDP), New York.

[4] Yang, L., (2014). An inventory of composite measures of human progress, Technical report, United Nations Development Programme Human Development Report Office.

[5] Saisana, M., & Tarantola, S. (2002). State-of-the-art report on current methodologies and practices for composite indicator development. European Commission, Joint Research Centre, Institute for the Protection and the Security of the Citizen, Technological and Economic Risk Management Unit, Ispra, Italy.

[6] Steenbergen, M., & Jones, B. (2002). Modeling Multilevel Data Structures. *American Journal of Political Science*, *46*(1), 218-237.

[7] Dedrick, R., Ferron, J., Hess, M., Hogarty, K., Kromrey, J., & Lang, T. et al. (2009). Multilevel Modeling: A Review of Methodological Issues and Applications. *Review Of Educational Research*, *79*(1), 69-102.

[8] Mazziotta, M., & Pareto, A. (2013). Methods For Constructing Composite Indices: One For All Or All For One? *Rivista Italiana Di Economia Demografia e Statistica*, *67*(02), 67-80.

[9] OECD. (2008). Handbook on constructing composite indicators: methodology and user guide. Paris.

[10] Sharma, S. (1996). *Applied multivariate techniques*. New York: J. Wiley.

[11] Fernando, M., Samita, S., & Abeynayake, R. (2012). Modified Factor Analysis to Construct Composite Indices: Illustration on Urbanization Index. *Tropical Agricultural Research*, *23*(4), 327.

[12] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: with applications in R.* New York: Spriger.

[13] Bartholomew, D. (2010). Analysis and Interpretation of Multivariate Data. *International Encyclopedia Of Education*, 12-17.

[14] De Leeuw, J., & Meijer, E. (2010). Handbook of multilevel analysis (pp. 1-75). New York: Springer.