Using Decision Trees to Improve Variable Selection for Building Composite Indicators

Adrian Oțoiu[1] | *Bucharest University of Economic Studies, Bucharest, Romania*
Emilia Țițan | *Bucharest University of Economic Studies, Bucharest, Romania*

Abstract

The established variable selection methods for building composite indicators have strong limitations with respect to the results obtained. Some of them focus on getting an index structure with a high alpha reliability and/or a high percentage of the total data variance explained. These methods are likely to omit variables with strong explanatory power, and lead to an unsatisfactory classification of countries. Decision trees can also be used in selecting variables that are the most relevant for building composite indicators. An example of variable selection for building a composite indicator, which compares results using Cronbach's coefficient alpha, factor analysis, and decision trees, shows that the latter method yields comparable, or better results. Using cluster analysis on the selected variables, we show that the decision tree variable shortlist has better discrimination power than those obtained with the other methods, even in the presence of outliers and missing values.

Keywords	JEL code
Composite indicators, decision trees, Cronbach's Alpha, factor analysis, indicator methodology, entrepreneurship indicators	*C43, I31, L26*

INTRODUCTION

The interest in designing and publishing composite indicators is not new. However, in the past few years, there was a strong increase in the number of composite indicators. According to The Economist (2014a), their number increased from just under 20 before 1994 to about 100 in 2000–2004, and to over 150 active indicators for the period 2010–2014. They cover just about any dimension of human activity, from human development and social progress to the state of the environment, education performance, ease of doing business, democracy, corruption, entrepreneurship, etc. There is an index for any social issue or public policy (The Economist, 2014b).

While the aim of these indicators is a laudable one, to aggregate several measures relevant to one domain in a simple, easy-to-understand index by governments, think-tanks and campaigners (The Economist, 2014b), it is very often the case that they are found wanting. In some cases they are misleading, leading to rankings that defy common sense and are disproved by empirical evidence

[1] Bucharest University of Economic Studies, Piata Romana 6, Bucharest, Romania. Corresponding author: e-mail: otoiu.adrian@gmail.com.

(Otoiu et al., 2014). In other cases, there are significant overlaps between the aims and contents of several indicators, e.g. Global Entrepreneurship Monitor, Global Competitiveness Index, and several other indicator frameworks (Pekka et al., 2013), Global Gender Gap, Gender Inequality Index, Index of Women's Power and Glass Ceiling Index for gender inequality (The Economist, 2014a), Human Development Index, Legatum Prosperity Index and Social Progress Index for human development and social progress (Otoiu et al., 2014), Climate Analysis Indicators Tool (CAIT) and Climate Change Performance Index (CCPI) for climate change (Bandura, 2008).

The information provided by the composite indicators is of great, and sometimes crucial, potential value. Nowadays, however, there is a great danger of misusing this information, and hiding essential evidence behind the global public attention/debate generated following (new) releases of composite indicators. In the age of artificial intelligence, this translates into a new level of accountability for the information provided by composite indicators with respect to their ability to provide an accurate picture of the state of affairs of a particular domain, as a whole and for specific issues they address, as big data analytics and the use and linkages between different datasets is now the main generator of knowledge (Rometty, 2018).

In spite of the popularity, widespread use and interest in composite indicators, it does not appear that competition among them has been effective in weeding out the good indices from the bad ones. Rather, it seems that some indices manage to survive despite having dodgy methodologies, which are in some cases not disclosed, due to a good management of celebrity endorsements and media coverage that creates headlines (The Economist, 2014b).

However, the key to having an index that is a meaningful measure of a domain is the use of a sound methodology that ensures it is an effective multidimensional gauge of that domain. This consists, in the first stages, in identifying variables relevant to the domain that represents the main focus of the index, which is in many cases followed by making a selection from them based on their statistical properties. Then, appropriate methods of weighting and aggregation would ensure that composite indexes and/or sub-indexes are built in a reliable way, that yield measures relevant to the multidimensional concepts that they are supposed to quantify.

While there is sufficient knowledge and expertise available in choosing the variables for constructing models, there are few techniques singled out as established methods for variable selection. Reference manuals and handbooks point of to both ensuring relevance, timeliness, availability and trustworthiness of the source variables (Hsu et al., 2013), while the main statistical methods used for selecting the component variables are mainly principal component analysis/factor analysis, Cronbach alpha coefficient, and cluster analysis (OECD, 2008). In other cases, correlation analysis was used to remove candidate variables that were similar in content and had similar correlation patterns with other variables (Otoiu et al., 2014).

Considering the fact that all methods used in selecting component variables for composite indicators are based on multivariate analysis, by taking into account their relative contribution in explaining their overall variation, this paper proposes a novel method for selecting variables using the decision tree method. In order to test our approach, we make a comparison between it, the Cronbach's alpha measure, and exploratory factor analysis for selecting variables used to construct two human progress indexes, the Human Development Index 2014 and the Legatum Prosperity Index 2014. We used then cluster analysis to assess the extent to which selected variables enable us to obtain a better cluster structure in terms of the ability to explain the total variability of the data and of the quality of clusters obtained, assessed using silhouette widths (Rousseeuw, 1987).

1 LITERATURE REVIEW

Selecting input variables is done primarily on the basis of their relevance to the construction of the index. This is based on a thorough knowledge of the field and the availability of the data. According to OECD (2008) the first step in constructing a composite indicator is "developing a theoretical framework" by defining the concept that will be measured by the indicator, determining the sub-groups corresponding to it in the case of conceptual multidimensionality (e.g. well-being), and identifying the selection criteria and methodology for component variables. The next step, which constitutes the focus of this paper, consists of selecting the variables that make up the index, or sub-indexes when there are subgroups, based on their strengths and their quality (OECD, 2008). Quality is often addressed in the methodological papers of several indexes, which shows that variables are selected based on their reliability, availability and timeliness (Porter and Stern, 2014; Legatum Institute, 2012).

OECD (2008) establishes that "an index is above all the sum of its parts" which should be kept in mind when building and assessing the outcomes of composite indicators. With respect to the specific properties that composite indicators should have, Paun (1983), cited by Pele (2008) establishes that "A composite indicator sensitive and anti-catastrophic is compensatory". This means that: 1) the improvement recorded in one component should be reflected in the improvement in the overall indicator Pele (2008), 2) it is not possible that a small change in one component effects a large change in the composite indicator. Both properties mean that a big change of one component is not accompanied by an opposite modification of another, so that the resulting values of the composite indicator will equal those obtained in the absence of this change (Pele, 2008).

Michalos et al. (2011) propose the following properties for an acceptable indicator: 1) relevant to the concerns of our main target audiences, 2) easy to understand, 3) reliable, valid, and sensitive to changes, 4) politically unbiased, 5) timely, easy to obtain, and periodically updated, 6) comparable across jurisdictions and groups, 7) objective or subjective, 8) positive or negative, 9) obtained through an open, transparent, and democratic consultative review process.

There are essentially two types of indicators, ones that are based on a relatively large number of variables e.g. Legatum Prosperity Index, Social Progress Index, and others that are based on only a few variables, such as the Human Development Index. While there are differences in the methodological approaches, most composite indicators that are reliable and enjoy a certain degree of prestige are based on a well-documented and scientifically sound methodology. A part of this methodology covers how component variables are selected. According to OECD (2008), the main statistical methods of selection are the following: principal component analysis/factor analysis, Cronbach alpha coefficient, and cluster analysis, which are widely used in practice to establish the shortlist of variables. A recent review paper (Gan et al., 2017) points out that the main analytical techniques used for constructing 96 sustainability indicators are principal component analysis/factor analysis and regression analysis, which both account for about 59% of all analytical techniques used. Of the total (analytical, opinion based, and equal weights) techniques used in weighting component variables of composite indicators and sub-indicators (weighting being a step in constructing aggregate indicators that follows variable selection), 11.46% employ principal component analysis/factor analysis, and another 6.25% employ regression method based on the use of a dependent variable relevant to the 'target', that is to the composite indicator or a sub-indicator (Gan et al., 2017).

Recently, some approaches have expanded and challenged the variable selection methods. One issue is the tradeoff between the choice of either a few variables, or a wide range of variables to capture the latent aspects of multidimensional concepts for which no specific variables to describe it are available (Foa and Tanner, 2012). Another one is the use of a "target measure" to help select the relevant variables for building a composite indicator. In this respect, the work of Abberger et al. (2018) shows that, even if established methods such as principal components and correlation analysis are used for variable selection, there is still the need to establish some reference variables that will guide construction and revision of composite indicators.

In this context, decision trees represent a predictive data mining technique whose primary use is to predict the outcome of one target variable based on the evolution of several explanatory variables.

Its results recommend it for the use of variable selection as the algorithm has the ability to discriminate between variables that have a significant explanatory power of the evolution of the target (dependent) variable. While there are other similar techniques which model the evolution of the dependent variables based on one or more independent variables (e.g. linear regression), decision trees are deemed to have the following advantages (Enachescu, 2009): ease of implementation and interpretation, robustness with respect to outliers and missing values, variable selection done taking into account interactions between variables.

Decision trees are also used in modelling the relationship between one target variable and several explanatory variables. In several fields, e.g. credit risk modelling and marketing, they are considered to be established techniques, which yield results comparable to other methods. Results by Hand and Henley (1997) show that decision tree methods yield comparable results to linear regression and logistic regression. Also, Siddiqui (2005) identifies decision trees as one of the key classification techniques used in statistical-based customer segmentation for credit scoring, and lists them as one of the methods used in building scorecards, along with logistic regression and neural network techniques. Customer segmentation using decision trees is popular in the telecom industry, used in predicting customer value (Weiss, 2005). Other applications of decision trees are found in medical sciences, in assessing the relative importance of variables identified as risk factors for major depressive disorder (Song and Lu, 2015) or for other diseases.

2 A BRIEF OVERVIEW OF THE VARIABLE SELECTION METHODS

The OECD handbook for constructing composite indicators (OECD, 2008) mentions three methods for selecting variables: factor analysis/principal components analysis (FA/PCA), scale reliability and cluster analysis.

FA/PCA emerges as a method of choice for selecting variables at it enables researchers to see the relationships between variables (OECD, 2008). Essentially, for variable selection, the exploratory factor analysis methods are used, which entail representing the observed variables as a function of parameters computed for unobserved factors (Preacher et al., 2013).

As Cooper (1983) points out, factor analysis comprises a set of multivariate statistical techniques, among which principal components analysis (PCA). While similar to FA in terms of results and use, PCA does not assume any relationship between the underlying structure of the variables (Cooper, 1983), and the components extracted (the equivalent of factors in FA) are, by design, orthogonal to each other. Due to this consideration, we have chosen FA as it may be better suited for exploring the relationships between variables based on their communalities (Cooper, 1983).

The scale reliability method measures the internal consistency of several variables used in building a single indicator (OECD, 2008). Its use for variable selection ensures that selected variables measure a single dimensional item (Nardo et al., 2005). The consistency is measured by the Cronbach coefficient alpha, reported in two forms, an unstandardized and a standardized form. Following the recommendation of Falk and Savalei (2011), we will use the standardized form, due to the fact that later stages of composite index construction will certainly require the use of a normalization technique that makes possible the aggregation of variables with different units of measurement into a single aggregate measure.

The third method, cluster analysis, is mostly a descriptive tool used to group countries (OECD, 2008) with the purpose to give some insights of the overall structure of the variables involved (OECD, 2008). It does not give a direct assessment of the contribution of each variable, and, in our opinion, its value mostly lies with checking whether results of variable selection done with other methods match the clusters obtained for shortlisted variables, and are compatible with index scores (Otoiu et al., 2014). Given these considerations we will use cluster analysis to compare the results obtained using other variable selection methods in terms of the quality of the cluster structures obtained.

3 USING DECISION TREES TO SELECT VARIABLES

Decision trees is an important technique used in data mining/knowledge discovery. It consists of classification of a target variable into a set of classes, based on the values of explanatory variables (Rokach and Maimon, 2014). This is performed using an algorithm known as recursive partitioning (Izenman, 2008), which is a step-by-step process by which a node is split, or not, into child nodes (Izenman, 2008) by asking a sequence of Boolean questions of type: is a value of a variable, on which the split is done, lower than a threshold value, or not? The process starts with a first, or root, node, and ends when an optimal solution is found based on criteria specific to the algorithm used (Izenman, 2008).

Among the competing algorithms used for building decision trees, we have chosen one of the most widely used, Classification and Regression Tree (CART), developed by Breiman et al. (1984) due to the fact that is one of the most commonly used, yielding results that are clear and easy to interpret.

Many software packages compute variable importance, which shows which variables control the classification process (Izenman, 2008). This is achieved through calculating the sum of the improvement scores for each explanatory variable, for all the nodes where it acts as a primary or surrogate splitting variable (Gebre-Sellasie et al., 2011; Thierneau and Atkinson, 2019). In the rpart package, which contains an implementation of CART in R, and will be used for variable selection later in this paper, variable importance is expressed as a percentage for each variable selected for tree construction by scaling the improvement sums for all variables to 100, and discarding those with proportions smaller than 1 (Thierneau and Atkinson, 2019).

The use of decision trees for variable selection is not as straightforward as the use of factor analysis and cluster analysis, as these techniques do not require the existence of a target variable and thus are ideal candidates for establishing the structure of the new variable, the resulting composite indicator or sub-index. The apparent difficulty of using this technique, given by the fact that it is a "supervised learning method" which requires the existence of a target variable (Bishop, 2007), can be overcome by the use of a target variable whose explanatory power is well-known, trusted, and relevant with respect to the goals and concept behind the developed composite indicator. This practice is used by the Legatum Institute (2013) in developing its Legatum Prosperity Index based on linear regressions of potential input variables on GDP per capita and overall self-reported life satisfaction, both considered to be the most reliable actual dimensions of human well-being. A similar approach is used by the Abberger et al. (2018) for improving the Swiss composite leading economic indicator by selecting variables based on their relationship with a reference, or target, variable.

4 CASE STUDY: USING DECISION TREES IN VARIABLE SELECTION

The feasibility and use of decision trees in selecting variables for constructing or revising a composite indicator will be shown in an application which attempts to select input variables for building a composite indicator of well-being. Data comes from Otoiu et al. (2014), where a validation of three of the most popular indexes of well-being and human progress is performed, namely the Human Development Index (HDI), Happy Planet Index (HPI), and Legatum Prosperity Index (LPI). Validation is based on a list of candidate variables deemed relevant for defining the multidimensional concept of well-being: 1) growth of CO^2 emissions (CO2GRW), 2) CO^2 emissions per capita (CO2PerC), 3) forest area as a share of total land area (Forest), 4) Gross National Income (GNI) per capita (GNIPC), 5) greenhouse gas emissions per capita (GHGPC), 6) Gini coefficient for income (GINI), 7) life expectancy at birth (LifeExpB), 8) natural resource depletion (NResDep), 9) mean years of schooling (YRSSch), 10) labor force participation rate for men (PRM), 11) labor force participation rate for women (PRF), 12) total labor force participation rate (PRT), 13) overall life satisfaction (SATISF), 14) share of fossil fuels in fuel consumption (ShareFF), 15) share of renewables in resource consumption (ShareRen), 16) unemployment rate (UE),17) urban pollution (Upoll), 18) well-being (WellB), computed as the arithmetic mean of individual responses to the Ladder of Life question in the Gallup World Poll.

In order to prove the validity and relative performance of decision trees used as a variable selection method, we will compare its results with those obtained using Cronbach's coefficient alpha and FA method. For the target variable used in the decision tree method, we chose GNIPC and WellB, as the best available proxies for the subjective and, respectively, objective well-being. These are the two major sides of the concept of well-being, described both in terms of material progress and an improvement in the living standards and conditions that will enable individuals to reach their goals given the opportunities available within a country at a certain time. A detailed explanation of the concept of well-being, which explains its bivalent nature, can be found in Otoiu et al. (2014), and the rationale for this approach can be found in the construction of the LPI (Legatum Institute, 2013).

All estimations are done using the R package **rpart** for decision trees, the **factanal** function for FA and the **reliability** function Cronbach's coefficient alpha, the latter two implemented in the RCommander graphical user interface. Results will be checked with the cluster analysis method Partition Around Medoids, implemented in the R "cluster" package. Compared to other clustering methods, this one produces graphs and diagnostics that enable an easy assessment of the quality of the cluster structure, both as a whole and for individual clusters.

Table 1 Results of decision trees estimation

Regression Tree Statistics

Target variable	CP	nsplit	rel error	xerror	xstd
GNIPC	0.597278	0	1	1.00662	0.145317
	0.153842	1	0.40272	0.56682	0.134489
	0.041019	2	0.24888	0.41913	0.111481
	0.04	3	0.20786	0.39637	0.088197
WellB	0.703880	0	1	1.00771	0.085633
	0.157171	1	0.29612	0.33884	0.033161
	0.048922	2	0.13895	0.17359	0.019016
		3	0.09003	0.12174	0.016763

Variable importance (percent)

Target variable	Explanatory Variables							
	CO2PerC	LifeExpB	YRSSch	PRM	WellB	SATISF	UE	ShareRen
GNIPC	24	19	15	13	13	12	4	1

Target variable	Explanatory Variables								
	SATISF	GNIPC	LifeExpB	CO2PerC	YRSSch	PRF	PRT	Forest	Upoll
WellB	35	19	18	13	8	3	2	1	1

Note: CP – complexity parameter, nsplit – number of splits, rel error – relative misclassification error, xerror – cross-validated error, xstd cross-validated standard deviation.
Source: Authors' calculation based on Otoiu et al. (2014) data

The results of the decision trees built using the two target variables are presented in Table 1. They indicate that decision tree structures obtained are fairly robust, with a goodness of fit (R^2) of about 79% for the GNIPC variable and, respectively, 91% for the WellB variable, computed from the misclassification errors for the last split (rel. error).

The resulting tree structures themselves were not particularly useful as they were too restrictive, with only three variables used in computing the solution for GNI per capita (CO2PerC, LifeExpB and UE) and only one for Well-being (SATISF, which is a strong proxy for well-being). However, 13 out of 20 variables reported as having a significant relative importance were retained and used to build a cluster structure that would discriminate countries based on their inferred level of well-being.

Variable selection done using the Cronbach's Alpha coefficients on the same variables has yielded the results shown in Table 2.

Table 2 Variable selection results using the scale reliability measure

Variable name	Alpha	Std.Alpha	R (item, total)
CO2PerC	0.0019	0.9277	0.7392
GNIPC	0.6876	0.9005	0.8226
LifeExpB	0.0016	0.9077	0.6737
SATISF	0.0026	0.8972	0.7285
WellB	0.0026	0.8992	0.7169
YRSSch	0.0024	0.9129	0.6522

Note: R (item, total) represents the correlation between one variable and the average behaviour of all variables.
Source: Authors' calculation based on Otoiu et al. (2014) data

Selection was done by deleting items for which R item scores, which compute the correlation between each item and the sum of the other items (Fox, 2012), is small. The final selection has a much lower number of variables, but the standardized alpha of 0.922 corresponding to the selected variables shown in Table 2, indicates a strong variable structure that can be used for calculating a composite indicator, fact confirmed by item scores well above 0.5, showing that each retained element has a significant contribution to the construction of the composite index.

Table 3 Factor loadings and diagnostic measures of factor analysis

Factor loadings

Variable	Factor1	Factor2	Factor3	Factor4
CO2PerC	0.269	0.196	0.927	0.157
Forest	0.106	0.282	–0.365	n/a
GNIPC	0.544	0.551	0.453	n/a
LifeExpB	0.555	0.548	0.221	n/a

Table 3 (continuation)

Factor loadings

Variable	Factor1	Factor2	Factor3	Factor4
SATISF	0.948	0.236	0.2	n/a
ShareFF	0.134	0.162	0.266	0.938
Upoll	–0.151	–0.552	0.112	n/a
WellB	0.934	0.263	0.21	n/a
Diagnostic measures				
SS loadings	2.576	1.756	1.34	1.121
Proportion of variance explained	0.286	0.195	0.149	0.125
Cumulative variance explained	0.286	0.481	0.63	0.755

Source: Authors' calculation based on Otoiu et al. (2014) data

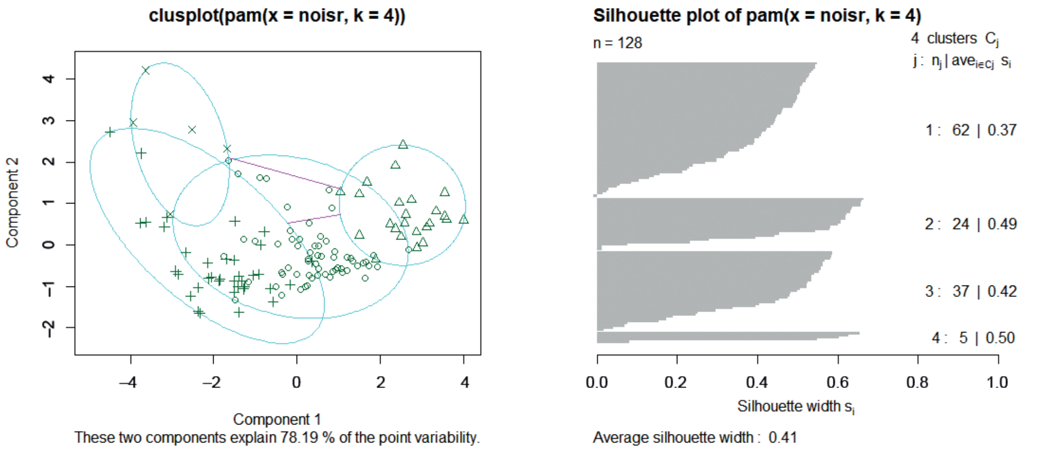
In achieving the factor solution presented in Table 3, we had to give up the GINI variable for which there was a high incidence of missing values. Results retain the variables for which factor loadings were higher than 0.5, as per the recommendation of the OECD composite indicators manual (OECD, 2008), and exclude the Forest variable which does not comply with this requirement. The solution presented still includes this variable as, by using it, we have obtained the best factor structure, which explains 75.5% of the total variability of the data set, and includes variables with significant factor loadings.

Further, in order to validate the results obtained using the three variable selection approaches, and show that they can be used to design a composite indicator that can classify countries based on their level of well-being, we use cluster analysis to compare the three sets of selected candidate variables in order to assess how well their total variability can be captured by cluster structures that classify countries into different groups based on their well-being.

The validation procedure is similar to the first part of the one done by Otoiu et al. (2014). An optimal cluster structure is obtained from selected input variables, with a high percentage of data variability being explained by the cluster structure (above 60%), silhouette plots averaging over 60% for the entire structure, and values of 50% or more, together with no misclassification for individual clusters (Otoiu et al., 2014). The cluster structures were obtained by eliminating observations with missing values. We considered this to be the right approach given the nature of the data and the exploratory aim of our approach. Using imputation techniques could have biased the results due to the fact that we deal with different countries whose unique characteristics may not be properly inferred through using a purely quantitative method. Due to this fact, in presenting our results, we have not compared the percentage of data variability explained by the cluster structure, as it would favor structures with fewer variables.

The optimal cluster structure obtained with Cronbach's Alpha coefficients, presented in Figure 1, shows that the selected variables, CO2PerC, LifeExB, SATISF, YRSSch, GNIPC, and WellB, manage to explain about 78% of the overall data variability. However, the cluster structure is fairly weak as average silhouette widths for the total cluster structure, and for three out of four individual clusters, is below 50%. Moreover, cluster membership is sizably unbalanced, with cluster 1 grouping almost half of the data, and cluster 4 having only 5 observations. The results are weak if the degree of misclassification is considered, shown

Figure 1 Optimal cluster structure with scale reliability variables

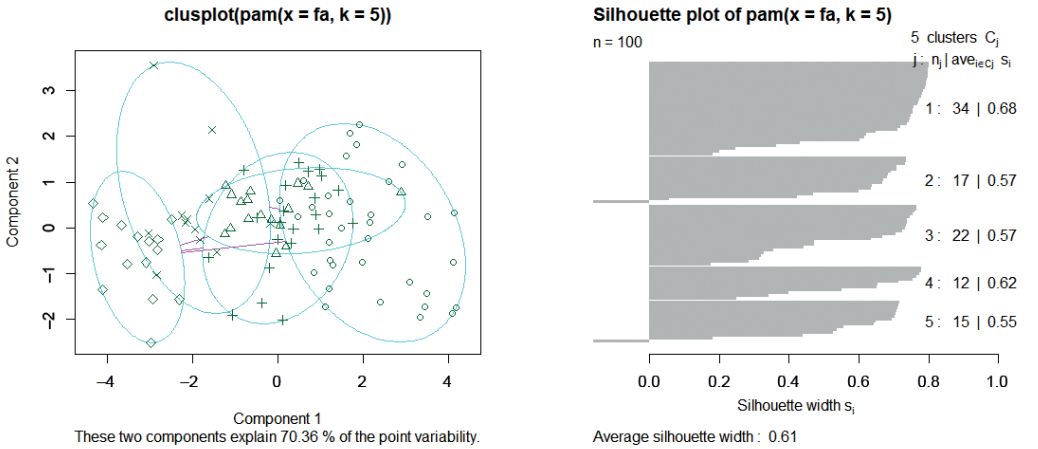


Source: Authors' calculation

as the part of the silhouette plots that extend below 0. As only marginal misclassification is observed, we conclude that, indeed, this structure is weak and cannot be used to achieve a satisfactory grouping of countries based on the selected variables.

For the factor analysis technique results, presented in Figure 2, show a balanced cluster structure, with an average silhouette width of 0.61, cluster individual widths above 0.55, and some sizable misclassification for cluster 2 and 5. Indeed, clustering based on 8 variables, CO2PerC, LifeExB, SATISF, YRSSch, GNIPC, WellB, Upoll, and ShareFF is able to achieve a fair discrimination of countries that would show their different levels of well-being in an appropriate way.

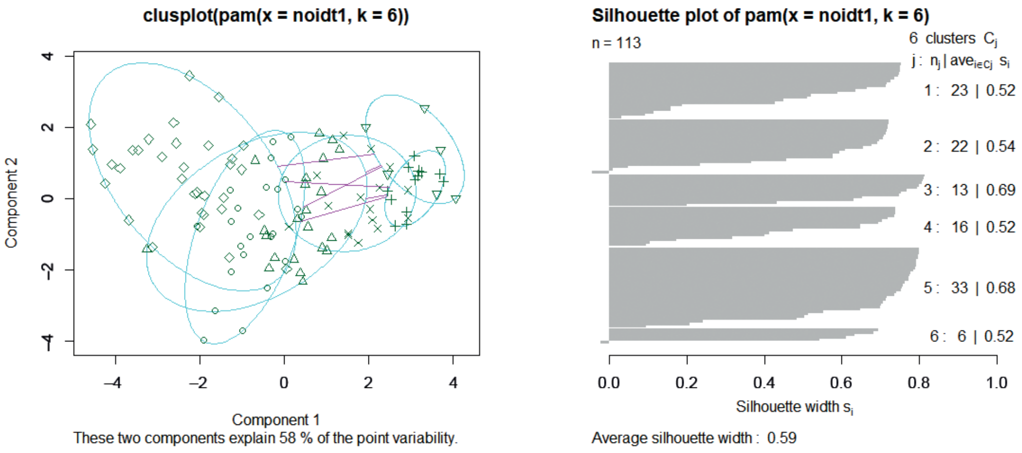
Figure 2 Optimal cluster structure with factor analysis variables



Source: Authors' calculation

The optimal cluster structure obtained for the variables selected with the decision tree algorithm, presented in Figure 3, has a slightly better performance than results obtained using factor analysis. The percentage of the overall variability explained is lower due to inclusion of 11 variables, CO2PerC, LifeExB, SATISF, YRSSch, GNIPC, WellB, Upoll, PRT, UE, Forest, and ShareRen. Albeit the average silhouette width is slightly worse (0.59 vs. 0.61) than for the structure obtained from variables selected using factor analysis, we see a very small of misclassification for cluster 2 and 6. Furthermore, all cluster-specific silhouette widths are above 50%, with rather high values of 68% or over for two clusters (3 and 5). Finally, for clusters 1 and 4, with the worst silhouette widths, no misclassification was observed. A comparison done with the six-cluster solution using variables selected with the FA method, presented in Appendix 1, shows a clear superiority of the solution using decision trees, with higher average silhouette width, strong misclassifications for 4 out of 6 clusters, and one cluster silhouette width below 0.5.

Figure 3 Optimal cluster structure with decision trees variables



Source: Authors' calculation

In sum, the decision tree method proved to be the most effective way of selecting candidate variables. The cluster structure obtained is able to discriminate six country groups with minor misclassification, obtaining clusters with a rather balanced number of elements. While some overall diagnostic measures may indicate that the cluster structure obtained using factor analysis for variable selection is better than the structure obtained using the variables selected with the rpart algorithm, interpretations should consider the decrease of average silhouette width which occurs when more variables are used, and the higher degree of misclassification observed for the former structure.

DISCUSSION AND CONCLUSIONS

Using decision trees to select variables for building composite indicators is a valid alternative to the established methods of variable selection. Its advantages lie in the fact that decision trees can work with virtually any type of variable, and that they are not very sensitive to outliers and missing values. Due to their features, variable selection is likely to be more complex and take into account the relevant features of the data set to a greater extent than when the other established methods are used.

The selection of relevant variables for defining well-being obtained with the decision tree algorithm, implemented with the rpart package, is better than the one which used Cronbach's alpha coefficients, and close to the one which employed factor analysis. The cluster structure obtained was markedly better, providing a higher number of country groups with little or no misclassification rate.

Several elements need to be taken into account when considering the use of decision trees as a variable selection method for composite indicators. The most important is the existence of variables which are representative proxies of the multidimensional concept that is analyzed. In our example, GNI per capita and Life satisfaction measures were used as the best available measures of well-being.

In the absence of an established proxy, building a naïve index with equal weights assigned to standardized/normalized candidate input variables, may provide an initial solution for the target variable to be used.

In some cases, decision trees can be extremely useful for designing parsimonious indexes, which employ only a few variables. A particular situation is relevant to the case of entrepreneurship indicators, when there are significant overlaps between variables (Pekka et al., 2013), and using decision trees can yield a better indicator structure that can clearly describe the multifaceted dimensionality of entrepreneurship.

Another example is HDI, which is calculated from 4 variables: Life expectancy at birth, Mean Years of Schooling, Expected Years of schooling, and GNI per capita. While this feature was not explored in this paper, further research may be able to assess the strengths of this method, and improve some of the composite indicators which use a large number of variables.

Another issue worth exploring is employing some rules for input variable selection using decision trees. It may be worth researching whether all variables selected as important are to be included in the development of a composite indicator, or a variable importance threshold should be established that would keep variables with large importance scores, and discard those with low importance, e.g. Urban pollution or Forest area in our case study.

ACKNOWLEDGEMENT

This paper was co-financed from the Human Capital Operational Program 2014–2020, project number: POCU/380/6/13/125245 No. 36482/23.5.2019 "Excellence in interdisciplinary PhD and post-PhD research, career alternatives through entrepreneurial initiative (EXCIA)", coordinator: The Bucharest University of Economic Studies.

References

ABBERGER, K., GRAFF, M., SILIVERSTOVS, B., STURM, J-E. Using rule-based updating procedures to improve the performance of composite indicators. *Economic Modelling*, 2018, 68, January, pp. 127–144.

BANDURA, R. A Survey of Composite Indices Measuring Country Performance: 2008 Update. *Studies Working Paper 2008–02*, United Nations Development Programme, Office of Development.

BISHOP, C. M. *Pattern recognition and machine learning.* New York: Springer-Verlag, 2006.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., STONE, C. J. *Classification and Regression Trees.* Wadsworth, Belmont, 1984.

COOPER, J. C. B. Factor Analysis: An Overview. *The American Statistician*, 1983, 37, 2, pp.141–147.

GAN, X. et al. When to use what: Methods for weighting and aggregating sustainability indicators. *Ecological Indicators*, 2017, 81, pp. 491–502.

GEBRE-SELASSIE, G., VAN BORTEL, W., LEGESSE, W., YEWHALAW D. Malaria and Water Resource Development. In: HUNTER, W. 3rd Ed. *Recent Advances and Issues in Environmental Science*, Oakville: Apple Academic Press, 2011.

ENĂCHESCU, D. *Data Mining: metode și aplicații. Editura Academiei Române*, 2009.

FALK, C. F. AND SAVALEI, V. The Relationship Between Unstandardized and Standardized Alpha, True Reliability, and the Underlying Measurement Model. *Journal of Personality Assessment*, 2011, 93, 5, pp. 445–453.

FOA, R. AND TANNER, J. C. Methodology of the Indices of Social Development. *ISD Working Paper Series 2012–04*.

FOX, J. *Documentation for package 'Rcmdr' version 1.8–4* [online]. 2008. [cit. 5.11.2015]. <http://artax.karlin.mff.cuni.cz/r-help/library/Rcmdr/html/00Index.html>.

IZENMAN, A. J. *Modern Multivariate Statistical Techniques.* New York: Springer, 2008.

HAND, D. J. AND HENLEY, W. E. Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society Series A*, 1997, 160(3), pp. 523–541.

HSU, A., JOHNSON, L. A., LLOYD, A. *Measuring Progress: A Practical Guide from the Developers of the Environmental Performance Index (EPI).* New Haven: Yale Center for Environmental Law & Policy, 2018.

LEGATUM INSTITUTE. *The 2012 Legatum Prosperity Index.* United Kingdom: The Legatum Institute, 2012.

LEGATUM INSTITUTE. *The Legatum Prosperity Index 2013. Methodology And Technical Appendix* [online]. United Kingdom: Legatum Institute, 2013. [cit. 2.11.2015]. <http://media.prosperity.com/2013/pdf/publications/Methodology_2013_FinalWEB.pdf>.

MICHALOS, A. C., SMALE, B., LABONTÉ, R., MUHARJARINE, N. T*he Canadian Index of Wellbeing. Technical Report 1.0* [online]. 2011. [cit. 6.2.2013]. <http://medcontent.metapress.com/index/A65RM03P4874243N.pdf>.

NARDO, M., SAISANA, M., SALTELLI, A., TARANTOLA, S. Tools for Composite Indicators Building. Joint Research Centre. *European Commission report EUR 21682 EN*, 2005.

OECD. H*andbook on Constructing Composite Indicators: Methodology and User Guide.* Paris: OECD, 2008.

OTOIU, A, TITAN, E, DUMITRESCU, R. Are the variables used in building composite indicators of well-being relevant? Validating composite indexes of well-being. *Ecological Indicators*, 2014, 46, pp. 575–585.

PAUN, G. An Impossibility Theorem for Indicators Aggregation. *Fuzzy Sets and Systems*, 1985, 9(2), pp. 205–210.

PEKKA, S., ZOLTAN, J. A., WUEBKER, R. Exploring country-level institutional arrangements on the rate and type of entrepreneurial activity. *Journal of Business Venturing*, 2013, 28(1), pp. 176–193.

PELE, D. T. About the Impossibility Theorem for Indicators Aggregation. *Journal of Applied Quantitative Methods*, 2009, 4(1), pp. 82–87.

PORTER, M. E. AND STERN, S. *Social Progress Index 2014* [online]. Social Progress imperative, 2014. [cit. 1.2.2018]. <https://www2.deloitte.com/content/dam/Deloitte/cr/Documents/public-sector/2014-Social-Progress-IndexRepIMP.pdf>.

ROUSSEEUW, P. J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics*, 1987, 20, pp. 53–65.

PREACHER, K. J., ZHANG, G., KIM, C., MELS, G. Choosing the optimal number of factors in exploratory factor analysis: a model selection perspective. *Multivariate Behavioral Research*, 2013, 48(1), pp. 28–56.

ROKACH, L. AND MAIMON, O. *Data mining with decision trees: theory and applications.* 2nd Ed. Singapore: World Scientific Publishing Co. Pte. Ltd., 2014.

ROMETTY, G. *We need a new era of data responsibility* [online]. World Economic Forum Annual Meeting, 2018. [cit. 2.2.2019]. <https://www.weforum.org/agenda/2018/01/new-era-data-responsibility>.

SIDDIQI, N. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring.* New York: Wiley and Sons, 2005.

SONG, Y. AND LU, Y. Decision tree methods: applications for classification and prediction. S*hanghai Archives of Psychiatry*, 2015, 27(2), pp. 130–135.

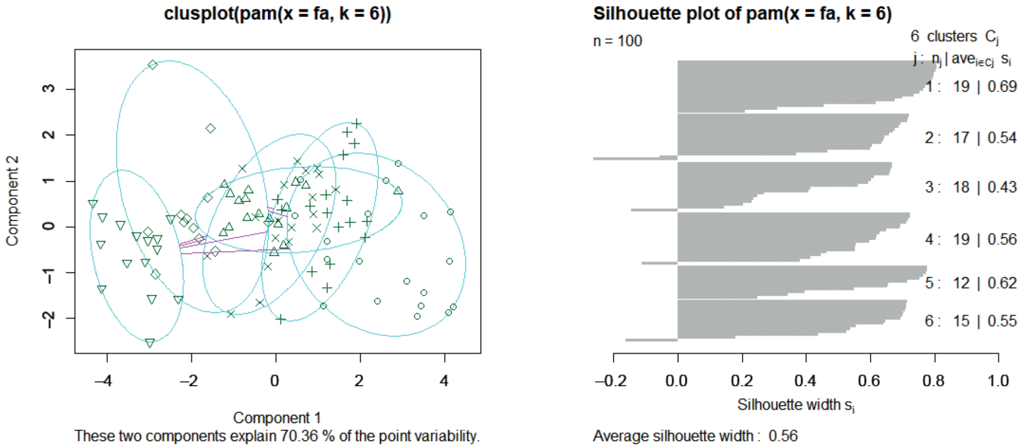
THE ECONOMIST. *Ranking the rankings.* November 8th 2014 print edition, 2014a.

THE ECONOMIST. *How to lie with indices.* November 8th 2014 print edition, 2014b.

THERNEAU, T. M. AND ATKINSON, E. J. *An Introduction to Recursive Partitioning Using the rpart Routine* [online]. 2019. [cit. 18.1.2020]. <https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf>.

ANNEX

Figure A1 The 6-cluster solution for variables selected with factor analysis



Source: Authors' calculation based on Otoiu et al. (2014) data