An ultrametric model to build a Composite Indicators system

Un modello ultrametrico per costruire un sistema di indicatori compositi

Carlo Cavicchia, Pasquale Sarnacchiaro, Maurizio Vichi and Giorgia Zaccaria

Abstract In the last years, the use of composite indicators has consistently increased, and the necessity to build model-based composite indicators with a strong methodological statistical approach becomes more and more important for reasons of trustworthiness. In this paper, we propose to build a composite indicators system able to measure different levels of relations among (group of) variables according to an ultrametric form which detects a hierarchical structure upon (group of) variables. Each dimension is measured as a specific composite indicator which reflects a subset of variables. In order to show its potential and applicability, the methodology is employed to analyze a dataset which contains variables about separated waste collection in Italy taking into consideration both its performance and its costs.

Abstract *Negli ultimi anni l'utilizzo di indicatori compositi è costantemente cresciuto, e la necessità di costruire degli indicatori compositi model-based con un forte approccio statistico è sempre più importante per motivi di fiducia. In questo articolo proponiamo di costruire un sistema di indicatori compositi che possa misurare diversi livelli di relazioni tra (gruppi di) variabili seguendo una forma ultrametrica che individui una gerarchia sulle (gruppi di) variabili. Al fine di mostrare il suo potenziale e la sua applicabilità, la metodologia è applicata per analizzare*

Carlo Cavicchia
Erasmus University Rotterdam, Rotterdam, The Netherlands
e-mail: cavicchia@ese.eur.nl

Pasquale Sarnacchiaro
University of Naples Federico II, Naples, Italy
e-mail: sarnacch@unina.it

Maurizio Vichi
University of Rome La Sapienza, Rome, Italy
e-mail: maurizio.vichi@uniroma1.it

Giorgia Zaccaria
University of Rome Unitelma Sapienza, Rome, Italy
e-mail: giorgia.zaccaria@unitelmasapienza.it

un dataset che contiene variabili riguardo la raccolta differenziata in Italia considerando sia le sue prestazioni che i suoi costi.

Key words: Latent variable model, Hierarchical model, Model-based, Latent concept, Statistical estimation

1 Introduction

Composite Indicators (CIs) are non-observable latent variables which consist of the aggregation of observed variables into a single non-observable index according to an underlying model for the multidimensional concepts [7, 8]. A CI is therefore a mathematical (weighted) combination of variables that generally is subject to several choices by the researcher [2]. CIs are able to summarize a big amount of information and for this specific feature they are very useful to measure multidimensional phenomena by potentially highlighting different levels of synthesis. However, the methods for CIs' construction are often criticized since they are not considered statistically rigorous or based on theories with solid foundations [6], thus, they might lead to misleading results and interpretations. This accounts for building CIs via a model-based approach.

In this paper, we propose to build a CIs system able to measure different levels of relations among (group of) observed variables according to an ultrametric structure [3, 4]. This structure allows describing multidimensional phenomena which are characterized by nested latent concepts having different levels of abstraction, from the most specific to the most general. In detail, internal consistent concepts are built and eventually aggregated from the most concordant ones to the most discordant. The proposal therefore detects a hierarchical structure upon variables.

In order to show its potential and applicability, the methodology is employed to analyze a dataset which contains variables about separated waste collection in Italy taking into consideration both its performance and its costs. This topic results crucial nowadays since many States are still land-filling huge amounts of municipal waste – the worst waste management option – despite the existence of better alternatives, and notwithstanding structural funds being available to finance better options. It is thus worth investigating how to measure the goodness and affordability of a waste management service via a system of CIs which assess each latent concept included in its definition. The number of information and statistics about waste management is larger and larger. For instance, Cavicchia, Sarnacchiaro and Vichi [1] detected which dimensions have an impact on the waste management in EU building a general composite indicator based on three specific composite indicators: recycling and circular economy performance, generation of recyclable waste, and private investments and innovation.

2 Methodology

CIs are able to reduce complex phenomena to a unique measure which results easier to interpret and might be used during the policy-making process. Notwithstanding their usefulness, it might be important to use both a set of specific indicators and a unique aggregated index. This means that a complex reality must be represented at different levels of abstraction and synthesis which might help to understand better the specific characterizations of the phenomenon that is being studied. The aggregated CI is the result of an entire hierarchy which starts from Q internal consistent latent concepts, in turn the hierarchy is provided by the ultrametric structure that reconstructs the main relationships among the variables. In other words, the hierarchy is composed of nested dimensions characterized by distinct levels of abstraction.

The different levels of the hierarchy are reconstructed thought four matrices: 1) a $p \times Q$ membership matrix $\mathbf{V}$, which represents the membership of each variable to a group where p is the number of the observed variables; 2) a diagonal matrix $\mathbf{S}^V$ of order Q, whose main diagonal represents the variance of each group; 3) a diagonal matrix $\mathbf{S}^W$ of order Q, whose main diagonal represents the covariance within each group; 4) a ultrametric matrix $\mathbf{S}^B$ of order Q, whose diagonal entries are set to zero and off-diagonal ones represent the hierarchical relationships among pairs of concepts. Whereas the CIs are built as the score vectors which best reconstruct the data matrix. The model-based approach which characterizes this hierarchy and this system of CIs guarantees to optimize an objective function in the least-squares framework.

The proposal extends the work by Cavicchia, Vichi and Zaccaria [5] by also reconstructing the covariance structure of the observed variables via an extended ultrametric covariance matrix [4]. However, the proposed method preserves the feature to obtain a reduced number of latent concepts which are quantified by maximizing the explained variance. These two goals are reach by minimizing a common objective function.

3 Application

The data used in this application about the separated waste collection are from different sources: Eurostat, Joint Research Centre (JRC) and Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA). It is worth observing that the observed variables used in this paper - namely, Cost of separated waste collection and transport, Cost of separated waste treatment and recycle, Organic waste collection, Paper waste collection, Glass waste collection, Metal waste collection, Plastic waste collection and Percentage of separated waste over the total waste - are regularly updated and free. Two variables represent the costs of the separated waste while the other six variables express the performance of it. In detail, we include in our analysis only the largest 40 Italian municipalities for comparability reason and we use the size population (i.e., number of inhabitants) and the size of waste produced (i.e., weight in

kilograms) to normalize the variables yet Percentage of separated waste over the total waste. This choice allows us to conduct two different analyses: one regarding the efficiency of the separated waste collection and one regarding the efficacy. A few steps of pre-processing are taken into consideration: the few missing data are Missing Completely at Random (MCAR) and therefore imputed by the K-nearest neighbors method by setting $K = 4$ and by using the Euclidean distance; and the variables are then standardized.

The motivation of this study lies on the assumption that it is crucial to combine the information from the costs and the performance to provide a support for Italian municipalities' actions and policies. The information from these two aspects, if measured separately, might be either misleading or limited. The research aims at searching other important latent concepts which might be present withing the main two, namely, costs and performance.

4 Conclusion

This paper provides a model-based approach to build a CIs system able to pinpoint a hierarchy and the quantification of the latent concepts which compose it. Furthermore, this study presents a useful tool to measure the separated waste collection in Italy together with its main aspects, by identifying the most important relationships among variables. The goal is to provide both a methodological contribution to the construction of CIs literature and a support for Italian municipalities' actions and policies.

References

1. Cavicchia, C., Sarnacchiaro, P. and Vichi, M.: A composite indicator for the waste management in the EU via hierarchical disjoint non-negative factor analysis. Socio-Economic Planning Sciences **73**, 100832 (2021)
2. Cavicchia, C. and Vichi, M.: Statistical Model-based Composite Indicators for tracking coherent policy conclusions. Social Indicators Research **156(2)**, 449-479 (2021)
3. Cavicchia, C., Vichi, M. and Zaccaria, G.: The ultrametric correlation matrix for modelling hierarchical latent concepts. Advances in Data Analysis and Classification **14(4)**, 837-853 (2020)
4. Cavicchia, C., Vichi, M. and Zaccaria, G.: Gaussian Mixture Model with an extended ultrametric covariance structure. Submitted. (2021)
5. Cavicchia, C., Vichi, M. and Zaccaria, G.: Hierarchical Disjoint Principal Component Analysis. Submitted. (2021)
6. Mazziotta, M. and Pareto, A.: Methods for constructing composite indices: One for all or all for one? Rivista Italiana di Economia Demografia e Statistica **67(2)**, 67-80 (2013)
7. Nardo, M., Saisana, M., Saltelli, A. and Tarantola, S.: Tools for Composite Indicators Building. European Commission (Join Research Centre, Ispra, Italy). Report EUR 21682. (2015)
8. Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A. and Giovannini, E.: Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing. OECD Statistics Working Papers 2005/3 (2005)