Computational Statistics
https://doi.org/10.1007/s00180-023-01333-9

ORIGINAL PAPER

A model-based ultrametric composite indicator for studying waste management in Italian municipalities

Carlo Cavicchia[1] · Pasquale Sarnacchiaro[2] · Maurizio Vichi[3] · Giorgia Zaccaria[4]

Received: 2 May 2022 / Accepted: 21 January 2023

Abstract

A Composite Indicator (CI) is a useful tool to synthesize information on a multidimensional phenomenon and make policy decisions. Multidimensional phenomena are often modeled by hierarchical latent structures that reconstruct relationships between variables. In this paper, we propose an exploratory, simultaneous model for building a hierarchical CI system to synthesize a multidimensional phenomenon and analyze its several facets. The proposal, called the Ultrametric Composite Indicator (UCI) model, reconstructs the hierarchical relationships among manifest variables detected by the correlation matrix via an extended ultrametric correlation matrix. The latter has the feature of being one-to-one associated with a hierarchy of latent concepts. Furthermore, the proposal introduces a test to unravel relevant dimensions in the hierarchy and retain statistically significant higher-level CIs. A simulation study is illustrated to compare the proposal with other existing methodologies. Finally, the UCI model is applied to study Italian municipalities' behavior toward waste management and to provide a tool to guide their councils in policy decisions.

Keywords Composite indicators · Hierarchical models · Ultrametricity · Waste management · External information

✉ Giorgia Zaccaria
giorgia.zaccaria@unimib.it

Carlo Cavicchia
cavicchia@ese.eur.nl

Pasquale Sarnacchiaro
sarnacch@unina.it

Maurizio Vichi
maurizio.vichi@uniroma1.it

[1] Erasmus University Rotterdam, Rotterdam, The Netherlands

[2] University of Naples Federico II, Naples, Italy

[3] University of Rome La Sapienza, Rome, Italy

[4] University of Milano-Bicocca, Milan, Italy

Published online: 16 March 2023

1 Introduction

Composite Indicators (CIs) have become increasingly relevant in the last twenty years as statistical tools for policy making. In fact, they are useful to convey and synthesize information on complex multidimensional phenomena that are not directly observable. As established by the Joint Research Centre and the Organization for Economic Co-operation and Development, a CI is an unobserved (latent) variable resulting from the aggregation of manifest variables into a single synthetic measure grounded in an underlying model of the multidimensional phenomenon under study (Nardo et al. 2005; OECD-JRC 2008). These complex phenomena are generally characterized by latent dimensions (concepts) ordered in a hierarchical structure that cannot be observable straightforwardly, and therefore, in turn, can be measured by CIs. Accordingly, two different types of CIs can be distinguished: the General Composite Indicator (GCI) to measure the multidimensional concept and the Specific Composite Indicators (SCIs) to represent its latent dimensions. This results in a CI system with an underlying hierarchical structure, where manifest variables are aggregated into first-level SCIs (specific dimensions), the latter into higher-level ones (broader dimensions) up to the GCI.

Some of the most severe criticisms of the use of CIs are related to the oversimplistic policy conclusions they can lead to and the normative approach to their construction, i.e., the fact that they are based on expert evaluation without any statistical assessment (Cavicchia and Vichi 2021; OECD-JRC 2008). However, both limitations can be overcome by building a model-based CI system. Indeed, even if a theoretical framework settled by a think tank can provide the interpretation of the phenomenon under study, the construction of a CI system via statistical models limits the researcher's arbitrary choices and connects it to the data via a mathematical formalization.

In the specialized literature, several methodologies have been proposed with the aim of modeling the multivariate data matrix or the covariance/ correlation matrix, to inspect the hierarchical relationships among manifest variables and detect latent dimensions and their quantification (Anderson and Rubin 1956; Cattell 1978; Wherry 1959; Schmid and Leiman 1957; Cavicchia and Vichi 2022, among others). These models were developed via a sequential approach, i.e., without optimizing an overall objective function, which can lead to inaccurate detection of the hierarchical relationships among manifest variables, or via a simultaneous approach, yet restricting the resulting hierarchy to a reduced number of levels. However, none of the existing methodologies builds a hierarchical structure over the manifest variables via a simultaneous approach, and by testing which levels of the hierarchy are not statistically significant so that to reduce their number and obtain a CI system representative of the researched multidimensional phenomenon. Therefore, this article aims to fill this gap by proposing a novel hierarchical methodology to build a CI system that considers all the hierarchical levels of the concept under study using a simultaneous model-based approach.

The proposal, called the Ultrametric Composite Indicator (UCI) model, unravels the hierarchical relationships between manifest variables by reconstructing the observed correlation matrix through an extended ultrametric one. The latter is a peculiar block matrix that has the features of being related one-to-one with a hierarchy of latent concepts and is represented by a tree structure, where the leaves correspond to the manifest variables, the internal nodes represent the SCIs, and the root identifies the GCI. Thus, an extended ultrametric correlation matrix results well-suited to model hierarchical relations among manifest variables and/or groups of them.

Notwithstanding being interesting from an interpretation point of view, not all internal nodes obtained as aggregation of those of lower level have to be necessarily retained in the hierarchy. In fact, if they are not statistically significant, some internal nodes corresponding to higher-level SCIs can be removed. In the UCI model, we introduce a test to evaluate the difference between two levels of the hierarchy engendered by the adopted ultrametric structure. The quantification of the CI system is based on the resulting hierarchy, where only statistically significant levels are maintained. It is worth underlining that the proposal is characterized by an overall objective function to optimize in order to obtain the optimal hierarchy in a simultaneous approach instead of a sequential and greedy manner. Moreover, the UCI model performs an exploratory analysis where only the number of first-level latent concepts is required beforehand. Differently from a confirmatory analysis, the exploratory one does not impose any relationships between manifest variables and first-level SCIs (or first-level and higher-level SCIs) by letting the data determine them, which can be extremely useful if a theoretical conceptualization of the phenomenon under study is not available or this is not empirically confirmed (see Cavicchia and Vichi 2021, for further details on the difference between exploratory and confirmatory analyses).

The performance of the UCI model is first evaluated through a simulation study on synthetic data, where it is compared with other existing methodologies for detecting hierarchical structures of variables. The proposal is then applied to the study of waste management in the 40 largest Italian municipalities by identifying its relevant latent dimensions. Waste management represents a multidimensional phenomenon that policy makers have highly considered in the last few decades (Heads of State and Government and High Representatives 2015; European Parliament and Council of the European Union 1999, 2018). To monitor waste collection and recycling in Europe, Eurostat collects indicators and statistics under the Waste Statistics Regulation (European Commission 2010), which can be used to build a waste management CI in Europe (Cavicchia et al. 2021). Starting from variables such that *Total costs of mixed waste management*, *Total costs of separated waste management* and *Percentage of separated waste over the total waste*, the proposal aims to pinpoint SCIs related to the quantities, performances, and costs of waste management and allows assessing the importance of each SCI in the construction of the GCI. Furthermore, the resulting SCIs and GCI are used to unravel different behaviors of Italian municipalities in waste disposal and treatment, as well as to determine the dimensions of waste management on which governments must focus to improve the performance of municipalities (i.e., increasing recycling practices and investing in separated waste). When studying the performance of Italian municipalities, it should be considered that several aspects can affect waste management and its dimensions. For example, tourism can have an impact on waste generation and collection (e.g., Matai 2015; Mateu-Sbert et al. 2013; Diaz-Farina et al. 2020). For this reason, we implement a further analysis considering aspects that affect waste management as external information and applying the UCI model to the data net of these effects.

The paper is organized as follows. In Sect. 2, the notation used throughout the paper is introduced and the definitions necessary to follow the specification of the model proposed here are provided. Section 3 thoroughly discusses the proposal in all its aspects (model specification, estimation, CI system definition and treatment of the external variable effect). The performance of the proposed model is illustrated in Sect. 4 both on synthetic and real data. A final discussion completes the article in Sect. 5.

2 Notation and background

For the convenience of the reader, the notation used in this paper is listed here.

n, p	Number of units and manifest variables, respectively
Q	Number of variable groups corresponding to the first-level SCIs over which the hierarchy is built
$\mathbf{X} = [x_{ij}]$	$(n \times p)$ data matrix
$\mathbf{R} = [r_{jl}]$	Data correlation matrix of order p, where r_{jl} is the correlation between the manifest variables j and l $(j, l = 1, \dots, p)$
$\mathbf{V} = [v_{jq}]$	$(p \times Q)$ membership matrix, where $v_{jq} = 1$ if the jth manifest variable belongs to the qth group; $v_{jq} = 0$ otherwise
$\mathbf{R}_{\mathrm{W}} = [{}_{W}r_{qq}]$	Diagonal matrix of order Q, whose diagonal entries represent the correlation within groups
$\mathbf{R}_{\mathrm{B}} = [{}_{B}r_{qh}]$	Matrix of order Q, whose off-diagonal entries represent the correlation between groups and diagonal ones are equal to zero
$\mathbf{E} = [e_{jl}]$	Error square matrix of order p
$\mathbf{Y}_{Q} = [y^{(Q)}_{iq}]$	$(n \times Q)$ score matrix of the first-level SCIs
$\mathbf{A}_{Q} = [a^{(Q)}_{jq}]$	$(p \times Q)$ sparse loading matrix, with a nonnull value per row representing the unique loading of each manifest variable on the corresponding first-level SCI. The position of each nonnull value per row is determined according to $\mathbf{V}$
$\mathbf{1}_p, \mathbf{1}_Q, \mathbf{I}_p$	Unitary vector of order p and Q, identity matrix of order p, respectively

Before going into the details of the model proposed to build a CI system in Sect. 3, we need to recall and introduce some definitions.

Definition 1 A matrix $\mathbf{U} = [u_{jl} \in \mathbb{R}_{\geq 0}]$ of order p is said to be ultrametric if

(i) $u_{jl} = u_{lj}$ for all $j, l = 1, \dots, p$ (symmetry);
(ii) $u_{jj} \geq \max\{u_{lj} : l = 1, \dots, p\}$ for all $j = 1, \dots, p$ (column pointwise diagonal dominance);

(iii) $u_{jl} \geq \min\{u_{ji}, u_{il}\}$, for all $i, j, l = 1, \dots, p$ (ultrametric inequality).

An ultrametric matrix has two main characteristics that make it suitable for building a hierarchy of composite indicators, starting with studying the relationships among manifest variables. These characteristics can be summarized as follows.

Remark 1 Every ultrametric matrix turns out to be *positive semidefinite* (psd) (Dellacherie et al. 2014, pp. 58-59).

Remark 2 An ultrametric matrix is associated one-to-one with a hierarchy of latent concepts (Cavicchia et al. 2020, 2022).

Remark 1 is essential if we analyze the relationships among the manifest variables through their correlations. In fact, a nonnegative correlation matrix of order p is an ultrametric matrix if (iii) holds, since (i) and (ii) are satisfied by definition. As we will see later in the paper, Remark 2 relates an ultrametric correlation matrix to a hierarchical structure. However, Definition 1 is based on the nonnegativity assumption, which can be very restrictive in several real applications. To include negative values and thus make the notion of ultrametricity more applicable in practice, the extension of Definition 1 is provided as follows.

Definition 2 A matrix $\mathbf{U} = [u_{jl} \in \mathbb{R}]$ of order p is said to be extended ultrametric if

(i) $u_{jl} = u_{lj}$ for all $j, l = 1, \dots, p$ (symmetry);
(ii.a) $u_{jj} \geq 0$ for $j = 1, \dots, p$ (nonnegativity of the diagonal);
(ii.b) $u_{jj} \geq \max\{|u_{lj}| : l = 1, \dots, p\}$ for $j = 1, \dots, p$ (column pointwise diagonal dominance);
(iii) $u_{jl} \geq \min\{u_{ji}, u_{il}\}$, for all $i, j, l = 1, \dots, p$ (ultrametric inequality).

It is worth noting that, if the nonnegativity assumption does not hold for the entire matrix, condition (ii.b) is not sufficient to guarantee the positive semidefiniteness of an extended ultrametric matrix, and thus to apply Definition 2 to a correlation matrix. To overcome this drawback, we request that if $\mathbf{U}$ is not psd, $\mathbf{U} = \mathbf{U} + a\mathbf{I}_p$, where a is the absolute value of the smallest eigenvalue of $\mathbf{U}$ (Cailliez 1983). This thus satisfies the positive semidefiniteness condition needed to apply the notion of ultrametricity to generic correlation matrices. In the next section, we introduce a new model-based approach for building a composite indicator system based on an extended ultrametric matrix.

3 The ultrametric composite indicator model

The Ultrametric Composite Indicator (UCI) model defines a hierarchy of composite indicators that starts with the study of the relationships among manifest variables and identifies broader dimensions associated with SCIs up to GCI. Therefore, we model the observed correlation matrix through an ultrametric structure to inspect the hierarchical relationships among manifest variables. This means that the UCI model reconstructs a correlation matrix $\mathbf{R} = [r_{jl} \in \mathbb{R}]$ of order p through an extended ultrametric correlation matrix $\mathbf{R}_{\mathrm{EU}}$, which is therefore psd, and an error square matrix $\mathbf{E}$ of the same order. Formally, the correlation matrix $\mathbf{R}$ of an $(n \times p)$ data matrix $\mathbf{X}$ is modeled by

$$\mathbf{R} = \mathbf{R}_{\mathrm{EU}} + \mathbf{E}, \tag{1}$$

where $\mathbf{R}_{\mathrm{EU}}$ detects the hierarchical structure of the manifest variables. Specifically, $\mathbf{R}_{\mathrm{EU}}$ is parameterized as follows

$$\mathbf{R}_{\mathrm{EU}} = \mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}' - \mathrm{diag}\left(\mathbf{V}\mathbf{R}_{\mathrm{W}}\mathbf{V}'\right) + \mathbf{V}\mathbf{R}_{\mathrm{B}}\mathbf{V}' + \mathbf{I}_p, \tag{2}$$

subject to constraints

$$\mathbf{V} = [v_{jq} \in \{0, 1\} \,:\, j = 1, \dots, p, q = 1, \dots, Q]; \tag{3}$$

$$\mathbf{V}\mathbf{1}_Q = \mathbf{1}_p \quad \text{i.e.} \quad \sum_{q=1}^{Q} v_{jq} = 1 \quad j = 1, \dots, p; \tag{4}$$

$$\mathbf{R}_{\mathrm{W}} = \mathrm{diag}([{}_W r_{11}, \dots, {}_W r_{QQ}]); \tag{5}$$

$$\begin{aligned}&\mathbf{R}_{\mathrm{B}} = \mathbf{R}'_{\mathrm{B}}, \mathrm{diag}(\mathbf{R}_{\mathrm{B}}) = \mathbf{0}, {}_B r_{qh} \geq \min\{{}_B r_{qs}, {}_B r_{hs}\}\ q, h, s = 1, \dots, Q,\\ &s \neq h \neq q;\end{aligned} \tag{6}$$

$$\min\{{}_W r_{qq} \,:\, q = 1, \dots, Q\} \geq \max\{{}_B r_{qh} \,:\, q, h = 1, \dots, Q, h \neq q\}; \tag{7}$$

Remark that diag(·) denotes the diagonal matrix whose diagonal elements are those of the parenthesized matrix. It can be easily proved that $\mathbf{R}_{\mathrm{EU}}$ is in agreement with Definition 2. In fact, it is symmetric since (5) and (6) hold; it is nonnegative on the diagonal and is column pointwise diagonally dominant since its diagonal corresponds to a unitary vector, that is, the diagonal of $\mathbf{I}_p$ in Eq. (2), whereas its off-diagonal elements vary between −1 and 1; lastly, it fits the ultrametric condition thanks to Eqs. (6)–(7). Moreover, if $\mathbf{R}_{\mathrm{EU}}$ is not psd, it must be rewritten as follows $\mathbf{R}_{\mathrm{EU}} = \mathrm{diag}(\widetilde{\mathbf{R}}_{\mathrm{EU}})^{-\frac{1}{2}}\, \widetilde{\mathbf{R}}_{\mathrm{EU}}\ \mathrm{diag}(\widetilde{\mathbf{R}}_{\mathrm{EU}})^{-\frac{1}{2}}$, where $\widetilde{\mathbf{R}}_{\mathrm{EU}} = \mathbf{R}_{\mathrm{EU}} + a\mathbf{I}_p$ and a is set to the absolute value of the smallest eigenvalue of $\mathbf{R}_{\mathrm{EU}}$.

The matrix $\mathbf{R}_{\mathrm{EU}}$ defined in Eq. (2) depends on three parameters: $\mathbf{V}$, which represents the membership matrix that defines the partition of the variables into Q groups ($Q \leq p$), each associated with a specific dimension, $\mathbf{R}_{\mathrm{W}}$ and $\mathbf{R}_{\mathrm{B}}$ that determine the characteristics of the groups. Specifically, $\mathbf{R}_{\mathrm{W}}$ is a diagonal matrix of order Q, whose diagonal entries represent the correlations within the variable groups, and $\mathbf{R}_{\mathrm{B}}$ is a matrix of order Q, whose off-diagonal elements represent the correlations between pairs of groups. Given the ultrametricity constraint (6), $\mathbf{R}_{\mathrm{B}}$ has a reduced number of different values that is at most $Q - 1$. By construction, $\mathbf{R}_{\mathrm{EU}}$ is then a $(2Q - 1)$-extended ultrametric correlation matrix since it has at most $2Q - 1$ different values, i.e., Q in $\mathbf{R}_{\mathrm{W}}$ and $(Q - 1)$ in $\mathbf{R}_{\mathrm{B}}$. Moreover, recalling Remark 2, it should be noted that $\mathbf{R}_{\mathrm{EU}}$ is one-to-one associated with a hierarchy of latent concepts. In detail, since each variable belongs to only one latent dimension, any triplet (i, j, l) of variables will surely fall into one of the following possible scenarios: (a) all elements of the triplet belong to a single group; (b) the elements of the triplet belong to two distinct groups; (c) all elements of the triplets belong to different groups. These three scenarios correspond to the following correlation triplets: $({}_{W}r_{qq}, {}_{W}r_{qq}, {}_{W}r_{qq})$, $({}_{W}r_{qq}, {}_{B}r_{qh}, {}_{B}r_{qh})$ and $({}_{B}r_{qh}, {}_{B}r_{qk}, {}_{B}r_{hk})$, respectively. Furthermore, all triplets verify the ultrametric inequality due to constraints (6) and (7). Thus, in $\mathbf{R}_{\mathrm{EU}}$ the Q values ${}_{W}r_{qq}$ $(q = 1, \dots, Q)$ correspond to the variable aggregations in groups defined by $\mathbf{V}$, while the other $Q - 1$ values ${}_{B}r_{qh}$ $(q, h = 1, \dots, Q,\ h \neq q)$ represent the aggregations in pairs of the Q variable groups. Therefore, $\mathbf{R}_{\mathrm{B}}$ defines the hierarchical structure of the Q variable groups considering its $Q - 1$ values in decreasing order. This gives rise to broader groups and corresponding dimensions lumped together from the most concordant to the least concordant.

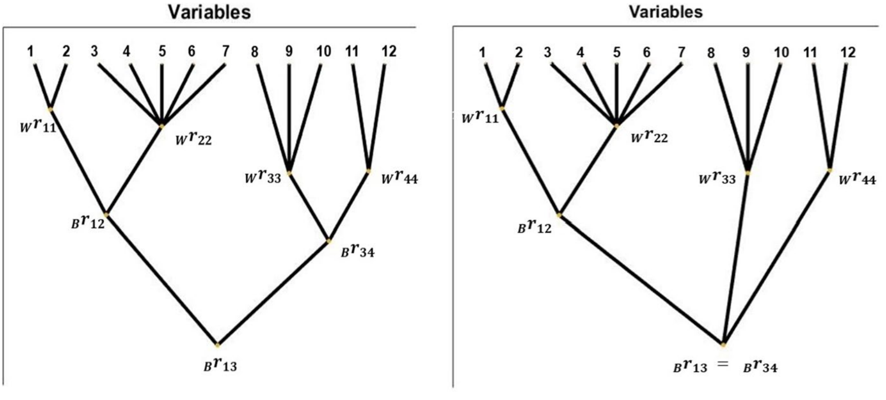
It has to be noted that constraint (7) allows us to guarantee that the variables belonging to the same group are more concordant among them than with the variables belonging to other groups, preserving the internal consistency of the Q variable groups. For this reason, a data preprocessing is recommendable. If a theory on the variable partition into Q groups exists, the UCI model can be applied in a semi-confirmatory approach, i.e., by constraining the membership of each variable to a specific group, where the polarity of the variables that are negatively related to the corresponding dimension is changed. $\mathbf{R}_{\mathrm{EU}}$ can also contain negative or zero values, other than positive ones. When this happens, the corresponding broader dimensions are defined by discordant or uncorrelated dimensions of lower levels, respectively.

An example of $\mathbf{R}_{\mathrm{EU}}$ and its parameters are provided in Fig. 1. Herein, four groups of variables can be detected: two variables are lumped together in the first group (first column of $\mathbf{V}$), five in the second group (second column of $\mathbf{V}$), three in the third group (third column of $\mathbf{V}$), and the last two in the last group (fourth column of $\mathbf{V}$). For simplicity reasons, the rows of the membership matrix $\mathbf{V}$ have been rearranged so that the variables belonging to the same group are contiguous. This variable partition corresponds to a block structure of $\mathbf{R}_{\mathrm{EU}}$, where the off-diagonal elements are equal to ${}_{W}r_{qq}$ $(q = 1, \dots, 4)$ if the corresponding two variables belong to the same group among the Q ones, or to ${}_{B}r_{qh}$ $(q, h = 1, \dots, 4, h \neq q)$ if the corresponding variables belong to two different groups and are lumped together further in the hierarchy. An example of the hierarchy corresponding to $\mathbf{R}_{\mathrm{EU}}$ is provided in Fig. 2a. Evidently, the order of aggregation between groups depends on the actual values of $\mathbf{R}_{\mathrm{B}}$ and therefore can be different from that shown in Fig. 2a.

$$\mathbf{V}=\begin{bmatrix}1&0&0&0\\1&0&0&0\\0&1&0&0\\0&1&0&0\\0&1&0&0\\0&1&0&0\\0&1&0&0\\0&0&1&0\\0&0&1&0\\0&0&1&0\\0&0&0&1\\0&0&0&1\end{bmatrix}\mathbf{R}_{\mathrm{W}}=\begin{bmatrix}{}_{W}r_{11}&0&0&0\\0&{}_{W}r_{22}&0&0\\0&0&{}_{W}r_{33}&0\\0&0&0&{}_{W}r_{44}\end{bmatrix}\mathbf{R}_{\mathrm{B}}=\begin{bmatrix}0&{}_{B}r_{12}&{}_{B}r_{13}&{}_{B}r_{13}\\{}_{B}r_{12}&0&{}_{B}r_{13}&{}_{B}r_{13}\\{}_{B}r_{13}&{}_{B}r_{13}&0&{}_{B}r_{34}\\{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{34}&0\end{bmatrix}$$

$$\mathbf{R}_{\mathrm{EU}}=\left[\begin{array}{cc|ccccc|ccc|cc}
1&{}_{W}r_{11}&{}_{B}r_{12}&{}_{B}r_{12}&{}_{B}r_{12}&{}_{B}r_{12}&{}_{B}r_{12}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}\\
{}_{W}r_{11}&1&{}_{B}r_{12}&{}_{B}r_{12}&{}_{B}r_{12}&{}_{B}r_{12}&{}_{B}r_{12}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}\\
\hline
{}_{B}r_{12}&{}_{B}r_{12}&1&{}_{W}r_{22}&{}_{W}r_{22}&{}_{W}r_{22}&{}_{W}r_{22}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}\\
{}_{B}r_{12}&{}_{B}r_{12}&{}_{W}r_{22}&1&{}_{W}r_{22}&{}_{W}r_{22}&{}_{W}r_{22}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}\\
{}_{B}r_{12}&{}_{B}r_{12}&{}_{W}r_{22}&{}_{W}r_{22}&1&{}_{W}r_{22}&{}_{W}r_{22}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}\\
{}_{B}r_{12}&{}_{B}r_{12}&{}_{W}r_{22}&{}_{W}r_{22}&{}_{W}r_{22}&1&{}_{W}r_{22}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}\\
{}_{B}r_{12}&{}_{B}r_{12}&{}_{W}r_{22}&{}_{W}r_{22}&{}_{W}r_{22}&{}_{W}r_{22}&1&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}\\
\hline
{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&1&{}_{W}r_{33}&{}_{W}r_{33}&{}_{B}r_{34}&{}_{B}r_{34}\\
{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{W}r_{33}&1&{}_{W}r_{33}&{}_{B}r_{34}&{}_{B}r_{34}\\
{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{W}r_{33}&{}_{W}r_{33}&1&{}_{B}r_{34}&{}_{B}r_{34}\\
\hline
{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{34}&{}_{B}r_{34}&{}_{B}r_{34}&1&{}_{W}r_{44}\\
{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{13}&{}_{B}r_{34}&{}_{B}r_{34}&{}_{B}r_{34}&{}_{W}r_{44}&1
\end{array}\right]$$

Fig. 1 Example of $\mathbf{R}_{\mathrm{EU}}$ and its parameters



(a) Complete hierarchy

(b) Hierarchy with statistically significant levels

Fig. 2 Hierarchy associated with $\mathbf{R}_{\mathrm{EU}}$ before (2a) and after (2b) the test

3.1 Estimation of the UCI model

Model (1) is estimated in a least-squares framework by fitting the closest extended ultrametric correlation matrix $\mathbf{R}_{\mathrm{EU}}$ to the correlation matrix $\mathbf{R}$. Hence, the optimization problem corresponds to minimizing the following loss function

$$F(\mathbf{R}_{\mathrm{W}}, \mathbf{R}_{\mathrm{B}}, \mathbf{V}) = \|\mathbf{R} - \mathbf{R}_{\mathrm{EU}}\|^2 \tag{8}$$

w.r.t. the parameters of $\mathbf{R}_{\mathrm{EU}}$ in Eq. (2) and subject to constraints (3)–(7). The details of the parameters' estimation are provided in Appendix A.

To find the parameter estimates $\widehat{\mathbf{R}}_{\mathrm{W}}$, $\widehat{\mathbf{R}}_{\mathrm{B}}$ and $\widehat{\mathbf{V}}$ that minimize Eq. (8), the least-squares estimation is performed via an algorithm that consists of the following steps: (0, initialization) a random partition $\widehat{\mathbf{V}}$ is generated from a Multinomial distribution in Q nonempty categories, each with equal probability, and the matrices reporting within and between groups correlations are computed accordingly; (1) the update of $\widehat{\mathbf{V}}$, subject to (3) and (4); (2) the update of $\widehat{\mathbf{R}}_{\mathrm{W}}$ and $\widehat{\mathbf{R}}_{\mathrm{B}}$ conditionally on the current configuration of $\widehat{\mathbf{V}}$ and subject to constraints (5)-(7); (3) the check on the positive semidefiniteness of the resulting extended ultrametric correlation matrix $\widehat{\mathbf{R}}_{\mathrm{EU}}$, which is obtained by substituting the results of Steps (1) and (2) into Eq. (2). The Steps from (1) to (3) are iteratively alternated and afterwards the loss function is computed. The latter decreases, or at least does not increase, at each iteration. The algorithm stops when the difference between the loss function in two sequential iterations is negligible, i.e., lower than an arbitrary small positive constant which is equal to 0.1^6 in our experiments. Because random initialization turns out to be prone to local optima, the algorithm is run several times (e.g., 100 in our experiments), starting from different random partitions of the variable space, to increase the chance to obtain a global minimum. However, the number of different solutions over the replications is limited, therefore, the algorithm results stable, and the presence of local optima does not result in an issue if the model runs 100 times.

A detailed and complete presentation of the algorithm for the estimation of the UCI model is provided in Appendix B, also including the test on the hierarchical levels produced by $\widehat{\mathbf{R}}_{\mathrm{EU}}$ and the computation of the SCIs and GCI on its significant levels, as discussed in the following two sections.

3.2 Test on the difference between two levels of the hierarchy

The hierarchy corresponding to $\mathbf{R}_{\mathrm{EU}}$ is composed of Q disjoint variable groups that identify the first hierarchical levels (the first four internal nodes that start at the top of Fig. 2a) and $Q - 1$ higher hierarchical levels that pinpoint their aggregations in pairs in broader groups, from the most concordant to the least concordant. As we will discuss in Sect. 3.3, the first Q internal nodes are crucial to unravel specific dimensions that account for the correlation among the manifest variables. Nonetheless, their aggregations – denoted into $\mathbf{R}_{\mathrm{B}}$ – could be irrelevant and the corresponding broader dimensions might result not statistically significant in the population. For this reason, it is pivotal to test whether the existence of all $Q - 1$ higher levels is statistically significant in order to retain the relevant dimensions in the hierarchy.

The test introduced herein is based upon that one proposed by Dunn and Clark (1969), and improved by Steiger (1980), for comparing correlations measured on the same individuals. We implement the test by analyzing the difference between the different values of $\mathbf{R}_{\mathrm{B}}$ that correspond to the aggregation between the variable groups. Starting from the last aggregation, which identifies the general concept (i.e., the root of the tree at the bottom of Fig. 2a), we test the difference between two subsequent values of $\mathbf{R}_{\mathrm{B}}$. Considering the example shown in Fig. 2a, the application of the aforementioned test is fulfilled by analyzing the difference between ${}_Br_{13}$ and ${}_Br_{12}$, and that one between ${}_Br_{13}$ and ${}_Br_{34}$.

In order to assess which out of the $Q-1$ higher levels are significant or can be discarded, the following hypothesis testing is performed

$$\begin{cases} \mathrm{H}_0 \ : \ {}_Br_{qh} - {}_Br_{ls} = 0 \\ \mathrm{H}_1 \ : \ {}_Br_{qh} - {}_Br_{ls} \neq 0 \end{cases}$$

where ${}_Br_{qh}$ and ${}_Br_{ls}$ are two correlations of $\mathbf{R}_{\mathrm{B}}$ that correspond to two sequential levels of the hierarchy. The test is performed by computing the following test statistic

$$Z = (z_{{}_B\hat{r}_{qh}} - z_{{}_B\hat{r}_{ls}})\sqrt{\frac{n-3}{2(1-\bar{s}_{qh,ls})}} \approx N(0,1), \tag{9}$$

where n is the sample size, $z_{{}_B\hat{r}_{qh}}$ and $z_{{}_B\hat{r}_{ls}}$ are the Fisher's z-transformations (Fisher 1921) of the sample estimators ${}_B\hat{r}_{qh}$ and ${}_B\hat{r}_{ls}$, respectively, and $\bar{s}_{qh,ls}$ is the sample estimator of the asymptotic covariance between $z_{{}_B\hat{r}_{qh}}$ and $z_{{}_B\hat{r}_{ls}}$ calculated using a pooled estimate of the correlation coefficients that are equal under the null hypothesis (see Steiger 1980, for further details). If the null hypothesis is rejected according to the test statistic in Eq. (9), then the hierarchical level (and the corresponding dimension) will be retained.[1]

The test is implemented from the last level of the hierarchy (that is, from the bottom to the top of Fig. 2a), since retention of the latter is fundamental for the construction of the GCI. Moreover, this choice is motivated by the goal of identifying latent dimensions, which are obtained by merging two dimensions of lower levels as much correlated as possible. Therefore, if the difference between two hierarchical subsequent levels is not statistically significant, no reason occurs to retain the lower level. Figure 2b displays an explanatory example of the effect of the test applied to the hierarchy obtained by the UCI model. The application of the test reveals only one statistically significant level in the hierarchy (${}_Br_{12}$), in addition to the last level corresponding to the GCI (${}_Br_{13}$); instead, the difference between ${}_Br_{13}$ and ${}_Br_{34}$ turns out to be not statistically significant and the corresponding hierarchical level is discarded. In this example, no other differences between hierarchical levels must be tested. The test stops when all the possible differences between two sequential hierarchical levels are tested, or equivalently when further tests on differences only include the first Q internal nodes.

[1] The rejection of H_0 occurs if $P(Z \geq |z_{obs}|) + P(Z \leq -|z_{obs}|) \leq \alpha$, where z_{obs} is the realization of the test statistic Z and α is the level of significance of the test set a priori.

3.3 Specific and General Composite Indicators scores

The test illustrated in Sect. 3.2 unravels which of the $Q-1$ higher levels resulting from $\widehat{\mathbf{R}}_{\mathrm{EU}}$ are statistically significant. According to its conclusion, the dimensions associated with the first Q internal nodes and the $H \leq Q-1$ statistically significant higher levels must be quantified. The quantification results into the definition of Q first-level[2] SCIs, $H-1$ SCIs of higher level associated with broader dimensions, and a GCI, that describes the multidimensional phenomenon of interest. The SCIs and GCI allow quantitatively evaluating the behaviors of units (e.g., countries) with respect to a dimension and/or a phenomenon and to make comparisons among them.

We can differentiate between the construction of first-level SCIs, higher-level SCIs and GCI as follows.

- *First-level SCIs*: the first Q SCIs, say $\mathbf{Y}_Q$, which correspond to the ones directly associated with manifest variables, are computed by selecting the principal component of maximum variance for each variable group. Therefore, for each $q=1,\dots,Q$, the manifest variables belonging to the qth group are considered to compute the principal component of maximum variance for the group (i.e., the qth column of $\mathbf{Y}_Q$). It should be noted that a reduced number of manifest variables is involved in the quantification of each first-level SCI since the Q variable groups are disjoint. For this reason, the loading matrix $\mathbf{A}_Q$ that contains the weight of each manifest variable on the corresponding component is sparse. Due to condition (4), each row of $\mathbf{A}_Q$ has only one nonnull element, which corresponds to the qth column of $\mathbf{V}_Q$ s.t. $v_{jq}=1, q \in \{1,\dots,Q\}$.
- *Higher-level SCIs and GCI*: for each higher hierarchical level, the corresponding SCI is computed by selecting the principal component of maximum variance for the SCIs of the lower level that compose it. The same holds for the GCI.

Looking at Fig. 2b, the first-level SCIs are those corresponding to the first four groups (from the top of the figure downward), each of which is calculated as the principal component of maximum variance for the manifest variables that define it (e.g., the second group is associated with variables 3, 4, 5, 6, 7); then the higher-level SCI, which is unique in this case, is obtained as the principal component of maximum variance resulting in a combination of the first-level SCIs of the groups 1 and 2; and finally, the GCI corresponding to the last aggregation is calculated as the principal component of maximum variance obtained considering the first-level SCIs associated with groups 3 and 4 and the higher-level SCI previously computed.

The choice of computing the principal components of maximum variance on the SCIs of lower levels is motivated by the idea to stress the importance to the hierarchy. Indeed, if each higher-level SCI were directly computed on the manifest variables, it would not take the levels of the hierarchy into account. Instead, the objective

[2] The internal nodes associated with the first Q SCIs have different levels of correlation, which correspond to the diagonal elements of $\mathbf{R}_{\mathrm{W}}$.

of the model is to obtain consistent and reliable first-level SCIs representing groups of highly positively correlated manifest variables and to build a hierarchy on them.

To define the variable groups and the corresponding first-level SCIs, Q must be determined. Indeed, the hierarchy obtained by $\widehat{\mathbf{R}}_{\mathrm{EU}}$ depends on the choice of Q, which identifies specific dimensions the multidimensional phenomenon is composed of. Q can be selected according to Kaiser's method (Kaiser 1960) and/or the unidimensionality (Cavicchia and Vichi 2021) of the first-level SCIs, among others. The latter corresponds to the evaluation of the second largest eigenvalue of the correlation sub-matrix of each variable group associated with a first-level SCI: if this is less than 1, then the corresponding SCI is unidimensional. Therefore, the optimal Q is chosen from 1 up to the value that corresponds to the first Q unidimensional first-level SCIs. The two aforementioned methods are used to choose the optimal number of first-level SCIs in the application presented in Sect. 4.2.

3.4 Cleaning composite indicators for external information

The researcher could be interested in considering additional information to build the CI system. In fact, the ranking of units based on the GCI (and SCIs) can be affected by some unit features that have not been considered in the analysis. In order to include external information, Takane and Shibayama (1991) proposed a decomposition of the original data into several components (see also Hunter and Takane 2002, for various applications of the proposed method). Specifically, we focus on the inclusion of auxiliary information on units, collected in the matrix $\mathbf{G}$ of dimension $(n \times r)$, where r is the number of external variables (i.e., external with respect to those of the original analysis). The model proposed by Takane and Shibayama (1991) is made up of two analyses: the external analysis and the internal analysis. In the first, the data matrix $\mathbf{X}$ is decomposed into a term that refers to what can be explained by $\mathbf{G}$, thus including the effect of external information, and another term that concerns what cannot be explained by $\mathbf{G}$, thus it is net of the effect of $\mathbf{G}$. In the latter, Principal Component Analysis (PCA, Pearson 1901; Hotelling 1933) is applied to some of the components or each component separately. In our case, the internal analysis is replaced by considering the UCI model.

We can summarize the procedure to include external information into the UCI model as follows.

External Analysis: The data matrix $\mathbf{X}$ is decomposed into two parts using the multivariate regression model, that is,

$$\mathbf{X} = \mathbf{G}\mathbf{C} + \mathbf{E}, \tag{10}$$

where $\widehat{\mathbf{C}} = (\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{X}$. By substituting $\widehat{\mathbf{C}}$ into Eq. (10), we obtain

$$\mathbf{X} = \mathbf{P}_G\mathbf{X} + \mathbf{Q}_G\mathbf{X},$$

where $\mathbf{P}_G = \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'$ and $\mathbf{Q}_G = \mathbf{I} - \mathbf{P}_G$ that, multiplied by $\mathbf{X}$, represent the original data with the inclusion of the effect of external information and net of this effect, respectively.

Internal Analysis: The correlation matrices of $\mathbf{P}_G\mathbf{X}$ and $\mathbf{Q}_G\mathbf{X}$ are computed, i.e., $\mathbf{R}^{(\mathbf{P}_G)}$ and $\mathbf{R}^{(\mathbf{Q}_G)}$, respectively. The UCI model could be applied on both separately.

In Sect. 4.2.3, we will focus on $\mathbf{R}^{(\mathbf{Q}_G)}$ in order to compute a CI system and evaluate differences in the GCI and SCI rankings of units net of the effect of additional information, that can affect the unit behavior towards the phenomenon under study.

4 Applications

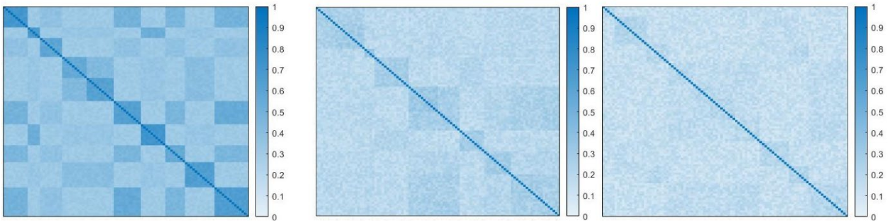
We carry out two analyses on synthetic and real data to assess the performance of the UCI model. In Sect. 4.1, we provide a simulation study where we compare our proposal with other existing methodologies. The UCI model is then applied to a real data set to study waste management in Italy in Sect. 4.2.

4.1 Synthetic data analysis

The performance of the UCI model in detecting hierarchical structures of variables is evaluated in comparison with the existing methodologies based upon sequential applications of PCA followed by oblique rotation methods, such that oblimin, quartimin, and geomin.

Two different scenarios are structured: one with a small scale correlation matrix and a small number of groups ($J = 30$ and $Q = 4$, respectively, *Scenario 1*), and another one with a large scale correlation matrix and a large number of groups ($J = 100$ and $Q = 10$, respectively, *Scenario 2*). The correlation matrices are generated according to Eq. (1). Specifically, the three parameters of $\mathbf{R}_{\mathrm{EU}}$ in Eq. (2) are obtained as follows: $\mathbf{V}$ is randomly generated from a Multinomial distribution in Q categories each with equal probability, where categories are not empty; the diagonal values of $\mathbf{R}_{\mathrm{W}}$ are generated as ${}_{W}r_{qq} = 0.85 + 0.1a$, where $a \sim N(0,1), q = 1,\dots,Q$, and the off-diagonal values of $\mathbf{R}_{\mathrm{B}}$ are set as ${}_{B}r_{qh} \in [0.4, 0.8]$, $q, h = 1,\dots,Q, h \neq q$, by keeping constant the difference between two sequential correlation coefficients and such that constraint (6) holds. In *Scenario 1*, the lower value of $\mathbf{R}_{\mathrm{B}}$ (the last aggregation) is set to negative. For each scenario, three levels of error are fixed: $\sigma_{\mathrm{E}}^{\mathrm{L}} = 0.1$ (low error), $\sigma_{\mathrm{E}}^{\mathrm{M}} = 0.5$ (medium error), and $\sigma_{\mathrm{E}}^{\mathrm{H}} = 0.9$ (high error). Error levels affect the generation of the error matrix $\mathbf{E}$, which is obtained by a uniform distribution in the interval $[0, \sigma_{\mathrm{E}}]$, symmetrized, and let it be positive semidefinite. The effect of the error level on the generation of the correlation matrix is shown in Fig. 3, where it can be seen that the variable groups and their hierarchical structure become less visible as the error level increases. The properties of the correlation matrix resulting from Eq. (1), i.e., the positive semidefiniteness and the appropriate range for its values are verified. For each scenario and error level, we generate 200 correlation matrices.

The comparison of the hierarchical structures pinpointed by our proposal and the competitors is carried out according to the Adjusted Rand Index (ARI, Hubert



(a) $\sigma_{\mathrm{E}}^{L} = 0.1$ (b) $\sigma_{\mathrm{E}}^{M} = 0.5$ (c) $\sigma_{\mathrm{E}}^{H} = 0.9$

Fig. 3 Example of heat maps of correlation matrices of order 100 produced with different levels of error (Scenario 2)

and Arabie 1985), that evaluates the similarity between the generated and the estimated partitions of variables. The ARI ranges between $-\infty$ and 1 (perfect agreement between the generated and the estimated membership matrix), and it is computed for each hierarchical level. For the UCI model the variable partitions in q, $q = Q - 1, \dots, 2$, groups are derived from the one in Q groups detected in $\mathbf{V}$ and the aggregations defined into $\mathbf{R}_{\mathrm{B}}$, whereas for the competitors they are obtained by assigning each variable (component) to the component (higher-order component) it loads more on in absolute term. It should be noted that the last aggregation is not taken into account, since it corresponds to the group containing all the variables. Moreover, the Mean Squared Error (MSE) of the parameters $\mathbf{R}_{\mathrm{W}}$ and $\mathbf{R}_{\mathrm{B}}$ is computed for all scenarios.

The results of the simulation study in terms of the mean of the ARI across the samples for the proposal and the competitors are provided in Table 1, whereas Table 2 shows the results of the MSE for the parameters of the UCI model. The proposed model turns out to have good results in terms of the mean of the ARI in all scenarios and for each level of error by outperforming the competitors. As expected, the performance of the UCI model, as well as that of competitors, decreases as the error level increases, as the latter tends to mask the hierarchical structure generated over the variables (Fig. 3). It is worthy to pinpoint that, differently from the UCI model, the mean of the ARI for the competitors usually declines as q lowers by stressing the difficulties in correctly detecting hierarchical relationships of variables with sequential models, even if they perfectly recover the variable partition in Q groups – as in the low error case. The UCI model also shows good performance in terms of the MSE of $\mathbf{R}_{\mathrm{W}}$ and $\mathbf{R}_{\mathrm{B}}$, as shown in Table 2.

4.2 Waste management in the largest Italian municipalities

In this section, the UCI model is applied to study waste management in the 40 largest Italian municipalities by identifying the latent dimensions and the corresponding SCIs that characterize it. The data set is presented in Sect. 4.2.1 and two analyses are performed. In the first one, the UCI model is implemented on the data set without considering any further information (Sect. 4.2.2); external variables are included in

Table 1 Mean of the ARI for the UCI model, PCA + Oblimin, PCA + Quartimin, PCA + Geomin for each hierarchical level

q	UCI model	PCA(O)	PCA(Q)	PCA(G)	UCI model	PCA(O)	PCA(Q)	PCA(G)	UCI model	PCA(O)	PCA(Q)	PCA(G)
	σ_E^L				σ_E^M				σ_E^H			
	Scenario 1											
2	1.00	0.51	0.51	0.24	1.00	0.96	0.96	0.57	1.00	0.80	0.80	0.49
3	1.00	0.99	0.99	0.98	1.00	0.98	0.98	0.94	0.99	0.71	0.71	0.56
4	1.00	1.00	1.00	1.00	0.95	0.88	0.88	0.88	0.89	0.73	0.73	0.71
	Scenario 2											
2	1.00	0.07	0.07	0.18	1.00	0.21	0.21	0.11	1.00	0.10	0.10	0.09
3	1.00	0.26	0.26	0.29	1.00	0.31	0.31	0.16	1.00	0.30	0.30	0.12
4	1.00	0.34	0.34	0.38	1.00	0.40	0.40	0.20	0.98	0.36	0.36	0.14
5	1.00	0.57	0.57	0.51	1.00	0.47	0.47	0.29	0.97	0.48	0.48	0.15
6	1.00	0.79	0.79	0.60	1.00	0.64	0.64	0.35	0.89	0.39	0.39	0.24
7	1.00	0.77	0.77	0.61	1.00	0.74	0.74	0.40	0.89	0.41	0.41	0.29
8	1.00	0.84	0.84	0.77	0.94	0.63	0.63	0.54	0.78	0.42	0.41	0.38
9	1.00	1.00	1.00	0.86	0.84	0.68	0.68	0.68	0.74	0.44	0.44	0.46
10	1.00	1.00	1.00	1.00	0.88	0.72	0.72	0.72	0.70	0.42	0.42	0.44

Table 2 MSE for the UCI model parameters

	Scenario 1			Scenario 2		
	$\sigma_{\mathrm{E}}^{\mathrm{L}}$	$\sigma_{\mathrm{E}}^{\mathrm{M}}$	$\sigma_{\mathrm{E}}^{\mathrm{H}}$	$\sigma_{\mathrm{E}}^{\mathrm{L}}$	$\sigma_{\mathrm{E}}^{\mathrm{M}}$	$\sigma_{\mathrm{E}}^{\mathrm{H}}$
MSE($\mathbf{R}_{\mathrm{W}}$)	0.00	0.03	0.06	0.01	0.03	0.04
MSE($\mathbf{R}_{\mathrm{B}}$)	0.01	0.05	0.09	0.01	0.08	0.10

Table 3 List of the 40 largest Italian municipalities

Ancona	Foggia	Parma	Salerno
Bari	Forli	Perugia	Sassari
Bergamo	Genova	Pescara	Taranto
Bologna	Livorno	Piacenza	Terni
Bolzano	Milano	Prato	Torino
Brescia	Modena	Ravenna	Trento
Cagliari	Monza	Reggio di Calabria	Trieste
Catania	Napoli	Reggio nell'Emilia	Venezia
Ferrara	Padova	Rimini	Verona
Firenze	Palermo	Roma	Vicenza

Table 4 List of the 13 manifest variables

ID	Label	Name	Measure
1	Costs of mixed waste collection and transport	CMWCT	€ per capita
2	Total costs of mixed waste management	TCMWM	€ per capita
3	Costs of separated waste collection and transport	CSWCT	€ per capita
4	Total costs of separated waste management	TCSWM	€ per capita
5	Percentage of costs of separated waste management over the total costs	PercCSW	%
6	Organic waste collection	OWC	kg per capita
7	Paper waste collection	PaWC	kg per capita
8	Glass waste collection	GWC	kg per capita
9	Wood waste collection	WWC	kg per capita
10	Metal waste collection	MeWC	kg per capita
11	Plastic waste collection	PlWC	kg per capita
12	Waste from electrical and electronic equipment	WEEE	kg per capita
13	Percentage of separated waste over the total waste	PercSW	%

the second analysis to take into account characteristics of Italian municipalities that could influence their performance in waste management (Sect. 4.2.3).

4.2.1 Data

The data used for waste management analysis were collected from Eurostat, Joint Research Centre and Istituto Superiore per la Protezione e la Ricerca Ambientale

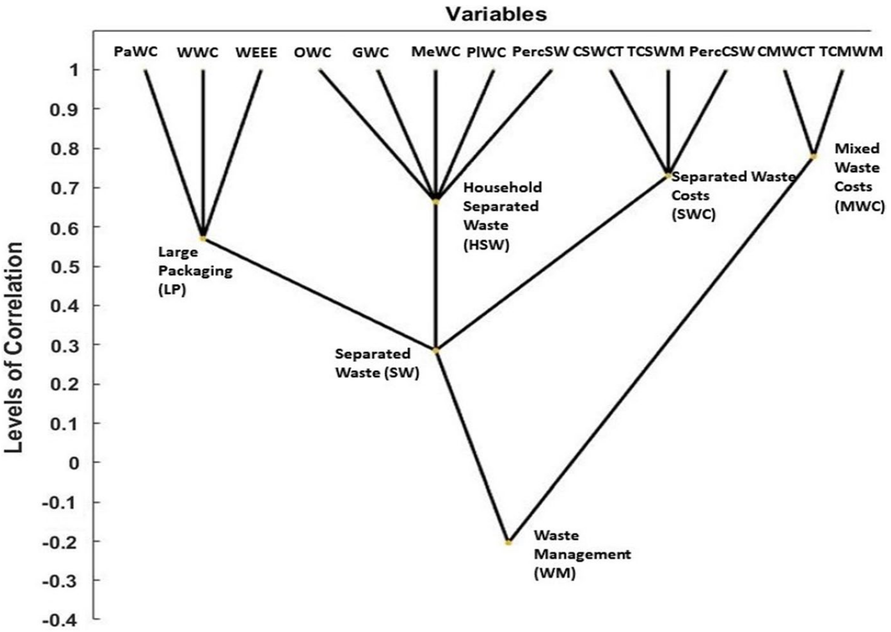


Fig. 4 Hierarchy resulting from the UCI model

for the 40 largest Italian municipalities (i.e., municipalities with more than 100.000 inhabitants) - 22 municipalities in the north, 8 in the center and 10 in the south and islands - at 2019 (Table 3). The data set consists of 13 manifest variables (Table 4) that are related to two main dimensions: costs (from 1 to 5) and quantities (from 6 to 13). For comparability reasons, the population size was used to normalized the manifest variables, when necessary. Few missing data occurred in the data set. They were Missing Completely At Random and were imputed via the K-nearest neighbors method by setting $K = 4$ and using the Euclidean distance. The manifest variables were standardized to z-score to eliminate the effect of different measurement units.

Other than the 13 manifest variables, 2 variables were included in the analysis as additional information for units: *Density*, which was computed as the ratio between the population size and the surface of the municipality (i.e., inhabitants per km^2), and *Touristic rate*, which was calculated as the total number of attendees in different accommodations over the population size of the municipality (i.e., total number of attendees per inhabitant). The municipalities with the highest density are Napoli, Milano, Torino, Palermo, Monza, Firenze, Pescara, Bergamo, Bologna, and Bari, while those with the highest touristic rate are Rimini, Venezia, Firenze, Ravenna, Roma, Verona, Trento, Milano, Bologna, and Padova (the *Density* and *Touristic rate* distributions are given in Fig. 1 of the Online Resource). The latter analysis allows us to take into account the influence of the density and touristic flows of a municipality on waste management, as we will see in Sect. 4.2.3.

Table 5 Results of the UCI model (loadings, unidimensionality, and Cronbach's α) in defining the dimensions of waste management

Variables / SCIs	MWC	SWC	HSW	LP	SW
1 - CMWCT	0.71				
2 - TCMWM	0.71				
3 - CSWCT		0.58			
4 - TCSWM		0.62			
5 - PercCSW		0.53			
6 - OWC			0.43		
7 - PaWC				0.58	
8 - GWC			0.42		
9 - WWC				0.61	
10 - MeWC			0.45		
11 - PlWC			0.43		
12 - WEEE				0.55	
13 - PercSW			0.50		
SW		0.41	0.75	0.52	
WM	−0.24				0.97
Unidimensionality	Yes	Yes	Yes	Yes	Yes
Cronbach's α	0.88	0.89	0.91	0.80	0.73

4.2.2 The UCI of waste management

Before applying the UCI model to the data set described in the previous section, the optimal number of first-level SCIs was selected. We determined Q according to the two different methods presented in Sect. 3.3: Kaiser's rule and unidimensionality. Both methods returned 4 as optimal Q.

The UCI model unravels one statistically significant higher level[3] in the hierarchy, in addition to those corresponding to the first-level SCIs and the GCI of Waste Management (WM), as shown in Fig. 4. As reported in Table 5, the first first-level SCI, that we called *Mixed Waste Costs* (MWC), is characterized by *Costs of mixed waste collection and transport* and *Total costs of mixed waste management*, which are both related to costs of mixed waste management. The second first-level SCI, named *Separated Waste Costs* (SWC), is defined by the three variables related to the costs of separated waste management, i.e., *Costs of separated waste collection and transport*, *Total costs of separated waste management* and *Percentage of costs of separated waste management over the total costs*. The third first-level SCI is characterized by *Organic waste collection*, *Glass waste collection*, *Metal waste collection*, *Plastic waste collection*, *Percentage of separated waste over the total waste*, and thus called *Household Separated Waste* (HSW); and the fourth first-level SCI is named *Large Packaging* as defined by *Paper waste collection*, *Wood waste collection* and *Waste from electrical and electronic equipment*. All first-level SCIs turn out to be unidimensional and reliable according to Cronbach's α (Cronbach 1951), since all are

[3] Significance level of the test: 0.05.

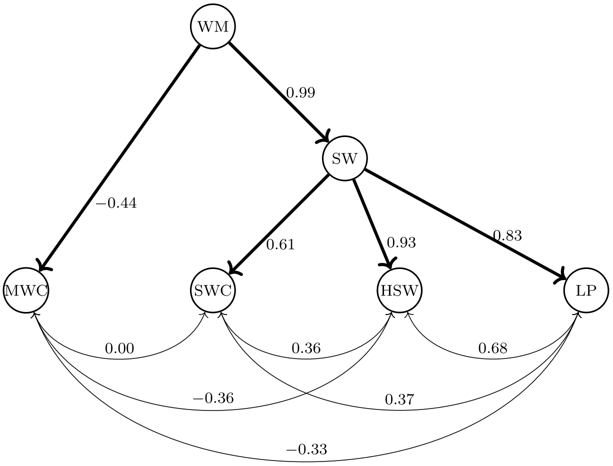


Fig. 5 Path diagram of the hierarchy resulting from the UCI model and representing the correlations between pairs of SCIs, and between SCIs and the GCI

greater than 0.7 (Table 5), which is considered as a threshold for acceptable value (Kline 2000). A higher-level SCI is obtained by merging SWC, HSW, and LP. This represents a latent dimension related to recycling (both costs and quantities), called *Separated Waste* (SW), which is mainly influenced by HSW and LP (see loadings in Table 5). Figure 5 detects positive relationships between SWC and HSW, and SWC and LP, that is, large amounts of separated waste progress at the same rate as the high costs of separated waste management, for example, for collection, transportation, etc.

The GCI of WM is then obtained by lumping together MWC (one of the first-level SCIs) and SW (the higher-level SCI), where the latter loads more on the GCI while the former has a negative relationship with it (Table 5). This means that the higher the quantities and costs of separated waste and the lower the costs of mixed waste, the better the waste management of a municipality. In fact, waste segregation is essential for proper recycling and avoids the use of landfills for waste disposal. Therefore, Italian municipalities that produce more separated waste and also invest more in it are those with the highest performance in waste management. It should be noted that the correlation between WM and SW (Fig. 5) is extremely high, and consequently we can evaluate the relationships between the GCI and the first-level SCIs of SWC, HSW and LP considering those between the latter and SW.

In Table 6, the rankings based on the GCI and SCIs are provided. They were obtained after normalizing the composite indicators by the Min-Max transformation. The rankings are substantially different, meaning that the behavior of each municipality can differ in the dimensions of WM. Taking into account the group of the best municipalities (reported in bold italic in Table 6), no municipality is in that group for all the SCIs and GCI, except for Ferrara, Rimini and Reggio nell'Emilia. If we consider the group of the worst municipalities (reported in italic in Table 6) on the GCI, we can notice that Catania is also in that group for all the SCIs, Genova as well except for MWC, whereas Palermo, Foggia and Taranto are in this group for two out of the four SCIs (HSW and LP). Roma, Venezia, Milano, Firenze, Napoli, Torino, Bologna, Verona, Bari – the cities classified by ISTAT as "large" – are in the

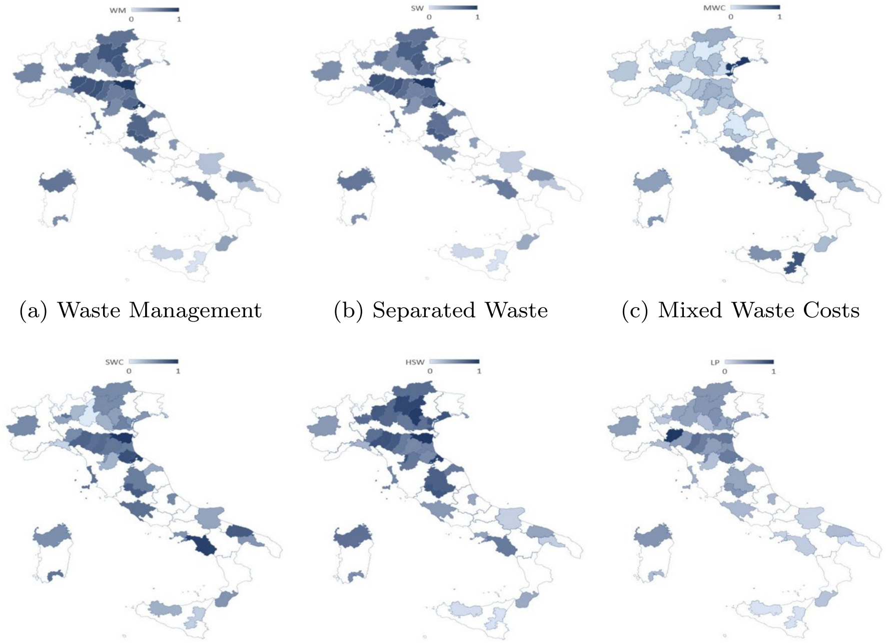
Table 6 Rankings based on normalized GCI and SCIs scores. Partition into groups according to thresholds: normalized score ≥ 0.60; normalized score ≥ 0.30 and < 0.60; normalized score < 0.30

WM	SW	MWC	SWC	HSW	LP
Ferrara	***Ferrara***	*Venezia*	***Ferrara***	***Ferrara***	***Piacenza***
Rimini	***Rimini***	*Catania*	***Salerno***	***Vicenza***	***Reggio nell'Emilia***
Reggio nell'Emilia	***Piacenza***	*Salerno*	***Rimini***	***Rimini***	***Ferrara***
Piacenza	***Reggio nell'Emilia***	**Roma**	***Forlì***	***Trento***	***Rimini***
Parma	***Vicenza***	**Palermo**	***Bari***	***Prato***	***Forlì***
Trento	***Terni***	**Cagliari**	***Terni***	***Parma***	**Modena**
Terni	***Parma***	**Foggia**	***Parma***	***Venezia***	**Prato**
Vicenza	***Trento***	**Sassari**	***Modena***	***Terni***	**Vicenza**
Prato	***Prato***	**Napoli**	***Reggio nell'Emilia***	***Bergamo***	**Sassari**
Perugia	***Modena***	**Bari**	***Livorno***	***Reggio nell'Emilia***	**Bolzano**
Modena	***Forlì***	**Bolzano**	***Roma***	***Perugia***	**Trento**
Forlì	***Venezia***	**Pescara**	***Cagliari***	***Brescia***	**Ravenna**
Bergamo	***Perugia***	**Rimini**	***Monza***	***Bolzano***	**Terni**
Padova	***Bolzano***	**Bologna**	***Bologna***	***Milano***	**Parma**
Bolzano	***Sassari***	***Piacenza***	***Ravenna***	***Piacenza***	**Torino**
Sassari	***Bergamo***	***Taranto***	***Perugia***	***Sassari***	**Bologna**
Monza	***Padova***	***Milano***	**Padova**	***Padova***	**Venezia**
Ravenna	**Salerno**	***Prato***	**Pescara**	***Monza***	**Brescia**
Livorno	**Ravenna**	***Reggio di Calabria***	**Torino**	***Firenze***	**Bergamo**
Venezia	**Monza**	***Verona***	**Bolzano**	***Verona***	**Padova**
Bologna	**Bologna**	***Ancona***	**Piacenza**	***Salerno***	**Perugia**
Salerno	**Livorno**	***Ravenna***	**Trento**	***Ancona***	**Trieste**
Brescia	**Milano**	***Modena***	**Sassari**	***Modena***	*Livorno*
Milano	**Cagliari**	***Vicenza***	**Napoli**	**Ravenna**	*Ancona*
Torino	**Ancona**	***Ferrara***	**Reggio di Calabria**	**Livorno**	*Pescara*
Ancona	**Torino**	***Trieste***	**Taranto**	**Bologna**	*Roma*
Firenze	**Brescia**	***Genova***	**Venezia**	**Forlì**	*Monza*

Table 6 (continued)

WM	SW	MWC	SWC	HSW	LP
Verona	**Roma**	***Reggio nell'Emilia***	**Prato**	**Cagliari**	*Cagliari*
Cagliari	**Verona**	***Forlì***	**Vicenza**	**Roma**	*Bari*
Bari	**Firenze**	***Terni***	**Milano**	**Torino**	*Genova*
Roma	**Bari**	***Livorno***	**Foggia**	**Trieste**	*Verona*
Pescara	**Pescara**	***Torino***	**Ancona**	**Napoli**	*Firenze*
Trieste	**Reggio di Calabria**	***Firenze***	**Palermo**	**Bari**	*Reggio di Calabria*
Reggio di Calabria	**Trieste**	***Monza***	**Verona**	**Pescara**	*Milano*
Napoli	**Napoli**	***Brescia***	**Firenze**	**Reggio di Calabria**	*Salerno*
Genova	*Genova*	***Bergamo***	*Bergamo*	*Genova*	*Napoli*
Taranto	*Foggia*	***Padova***	*Trieste*	*Taranto*	*Foggia*
Foggia	*Taranto*	***Parma***	*Catania*	*Foggia*	*Catania*
Palermo	*Palermo*	***Perugia***	*Genova*	*Palermo*	*Palermo*
Catania	*Catania*	***Trento***	*Brescia*	*Catania*	*Taranto*

The emphasis (italic, bold and bold italic) used for the ranking of MWC are reversed w.r.t. the others since the loading of this SCI on the GCI is negative



(a) Waste Management (b) Separated Waste (c) Mixed Waste Costs

(d) Separated Waste Costs (e) Household Separated Waste (f) Large Packaging

Fig. 6 Normalized GCI and SCIs scores for the 40 largest Italian municipalities

intermediate municipality group for the GCI. Other "large" cities such as Genova, Palermo, and Catania behave differently across the SCIs. For instance, Roma is in the group of the intermediate municipalities for MWC and HSW, in the group of the best municipalities for SWC and in the group of the worst municipalities for LP. Although, generally speaking, the smaller the quantity the better is in terms of waste, it has to be noted that *Percentage of separated waste over the total waste* has the highest loading on HSW. For this reason, we can state that the different position of Roma in the rankings of SWC and HSW could be due to an investment of this municipality on separated waste which does not still correspond to a high level of separate waste collection in terms of quantities.

The territorial distribution of the normalized scores of the GCI and the SCIs is represented in Fig. 6. For readability reasons, the map of Italy displays provinces instead of municipalities the data refer to; however, each municipality represents the main city of the corresponding province. The northern municipalities show to have a higher WM performance than the southern ones (Fig. 6a and Fig. 2 in Online Resource), which reflects the better behavior in separated waste, and, in particular, separated waste collection (Fig. 6e). It is noteworthy that the northern municipalities are also those with the lowest values of MWC (Fig. 6c), whereas LP in Fig. 6f shows values lower than those of the other SCIs in Italy. The latter may be due to the fact that the variable that loads more on LP is *Wood waste collection* (Table 5), whose collection also depends on specific characteristics of the municipalities, e.g., the presence of green areas.

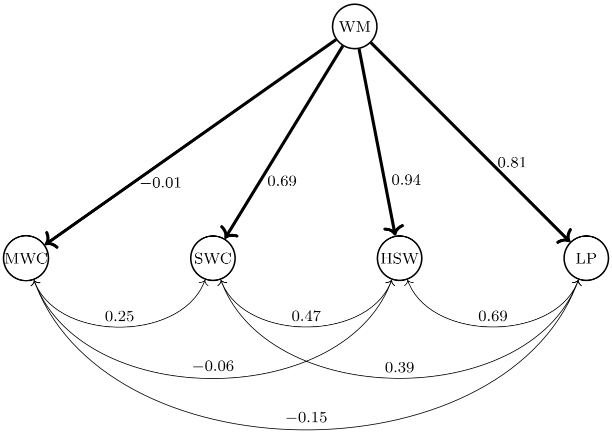


Fig. 7 Path diagram of the hierarchy resulting from the UCI model net of the effect of external information, and representing the correlations between pairs of SCIs, and between SCIs and the GCI

However, several features of the municipalities can affect their waste management. In fact, if we consider the 10 municipalities with the highest density (see Sect. 4.2.1), 7 are in the group of the intermediate municipalities for WM and 1 into that of the worst municipalities for WM (i.e., Palermo), whereas 6 out of the 10 municipalities with the highest touristic rate (see Sect. 4.2.1) are in the intermediate group of the WM ranking (Table 6).

In the next section, we analyze the UCI model applied on the data set net of the effect of *Density* and *Touristic rate*, which can affect, and make more difficult, the municipalities' waste management.

4.2.3 Influence of external variables

As introduced in Sect. 3.4, we considered the effect of external variables which can affect the behavior of the municipalities in waste management. In this case, the matrix $\mathbf{G}$ consists of the variables *Density* and *Touristic rate* measured in the 40 largest Italian municipalities. The goal of this analysis is to evaluate WM net of the effect of the *Density* and *Touristic rate* and to pinpoint differences in its ranking. Therefore, we focus on $\mathbf{Q}_G\mathbf{X}$. To compare the results, we fixed the membership of the 13 variables with the corresponding first-level SCI, according to the partition obtained in Sect. 4.2.2, and we let the UCI model identify the hierarchy and its statistically significant levels. Indeed, an important aspect of the UCI model is that it provides the possibility to fix some (or all) relationships between manifest variables and first-level SCIs in a semi-confirmatory approach when a theoretical framework on the phenomenon under study is known a priori or a previous analysis has already been carried out. The comparison can provide interesting information on differences among municipalities generated by external effects to the mere analyzed phenomenon. We thus implemented a semi-confirmatory approach for the UCI model, where only the first-level SCIs are fixed, as well as their number ($Q = 4$).

In this case, the UCI model does not pinpoint higher-level SCIs. Thus, only two levels exist in the hierarchy: one corresponding to the fixed first-level SCIs, and the other one to the GCI of WM. Looking at Fig. 7, it can be highlighted that the three first-level SCIs related to separated waste remain the most important in the definition of waste management, even if the loading of LP is reduced to 0.48, while that of SWC increases to 0.45, w.r.t. the same obtained without considering external information. The relationship between the GCI and the first-level SCI that is most affected by the removal of the effect of external information is with MWC. Indeed, its loading is reduced to -0.01 by omitting its impact in the definition of WM. It must be considered that both density and tourism have an impact on mixed waste. Specifically, density affects the production of mixed waste, as higher density limits the possibility of implementing door-to-door recycling collection due to smaller spaces. Furthermore, tourism waste is also mainly characterized by mixed waste and is therefore associated with higher costs. The tourist destinations often correspond to the cities' historic centers which are usually pedestrianized or restricted traffic zones. In the latter, mixed waste costs significantly increase because of the need to use vehicles of reduced dimensions, whose operating cost is higher than that of standard vehicles, and the higher presence of mixed waste bins.

Rankings based on the normalized scores of WM and first-level SCIs net of the effect of external information are shown in Table 7 and 8, respectively. Large cities such as Milano, Torino, Napoli, Venezia, Firenze, and Bologna, having the highest values for one or both external variables and being in the group of intermediate municipalities for WM in the previous analysis, belong to the group of the best municipalities for WM after removing the effect of *Density* and *Touristic rate*. This result supports the hypothesis that the density of a municipality and the flows of tourists make waste management more difficult, as well as waste separation, regardless of the territorial distribution of the municipalities (see also Fig. 3 of the Online Resource). On the contrary, the bottom end of Table 7, that is, the group of the worst municipalities, remains substantially unchanged. Moreover, considering separated waste (costs and quantities), Napoli is in the group of the intermediate municipalities for SWC and HSW, and in the group of the worse municipalities for LP if no external information is considered, whereas if the latter is treated in the analysis Napoli belongs to the group of the best municipalities for SWC and HSW, and the group of the intermediate municipalities for LP.

5 Conclusions

In this paper, we propose the UCI model to reconstruct the main hierarchical relationships among the manifest variables, which are represented by the correlation matrix. Distinct to the existing hierarchical methods, the proposal is simultaneous and minimizes an overall objective function for obtaining the hierarchical solution. To minimize the least-squares loss function, we present a block-coordinate descent algorithm. Moreover, the UCI model is characterized by the introduction of a statistical test for the hierarchical levels to consider into the hierarchy. The test leads to a further reduction in the number of CIs to include in the model by building a parsimonious CI system for the phenomenon studied.

Notwithstanding the fact that the model selection problems are addressed in the paper by providing indications on the appropriate selected number of first-level SCIs, it remains for future studies to consider other information criteria useful for such model selection.

The proposal has several applications in different fields, for example, to study climate change and its dimensions, to build a model-based CI system to track the Sustainable

Table 7 Ranking based on the normalized scores of WM net of the effect of external information on municipalities, compared to the ranking based on WM. Partition into groups according to the thresholds: normalized score ≥ 0.60; normalized score ≥ 0.30 and < 0.60; normalized score < 0.30

WM	WM net of the external variable effect
Ferrara	***Rimini***
Rimini	***Milano***
Reggio nell'Emilia	***Torino***
Piacenza	***Ferrara***
Trento	***Napoli***
Parma	***Venezia***
Vicenza	***Vicenza***
Terni	***Piacenza***
Prato	***Prato***
Perugia	***Monza***
Modena	***Reggio nell'Emilia***
Forlì	***Firenze***
Bergamo	***Bergamo***
Padova	***Trento***
Bolzano	***Parma***
Sassari	***Bolzano***
Monza	***Terni***
Ravenna	***Bologna***
Venezia	***Padova***
Livorno	**Salerno**
Bologna	**Modena**
Brescia	**Forlì**
Salerno	**Perugia**
Milano	**Roma**
Torino	**Bari**
Ancona	**Pescara**
Firenze	**Livorno**
Verona	**Ravenna**
Cagliari	**Brescia**
Bari	**Cagliari**
Roma	**Sassari**
Pescara	**Verona**
Trieste	**Ancona**
Reggio di Calabria	**Trieste**
Napoli	*Genova*
Genova	*Palermo*
Taranto	*Reggio di Calabria*
Foggia	*Taranto*
Palermo	*Foggia*
Catania	*Catania*

Table 8 Rankings based on the normalized scores of the four SCIs net of the effect of external information on municipalities. Partition into groups according to the thresholds: normalized score ≥ 0.60; normalized score ≥ 0.30 and < 0.60; normalized score < 0.30

MWC	SWC	HSW	LP
Venezia	***Rimini***	***Milano***	***Piacenza***
Catania	***Salerno***	***Rimini***	***Torino***
Salerno	***Napoli***	***Vicenza***	***Rimini***
Napoli	***Ferrara***	***Bergamo***	***Reggio nell'Emilia***
Palermo	***Torino***	***Ferrara***	***Ferrara***
Roma	***Bari***	***Napoli***	***Modena***
Milano	***Milano***	***Prato***	**Prato**
Rimini	***Forlì***	***Venezia***	**Milano**
Cagliari	***Monza***	***Trento***	**Forlì**
Bari	***Roma***	***Firenze***	**Bolzano**
Torino	***Bologna***	***Torino***	**Bologna**
Pescara	**Venezia**	***Monza***	**Vicenza**
Bologna	**Terni**	***Parma***	**Napoli**
Bolzano	**Parma**	***Brescia***	**Bergamo**
Foggia	**Modena**	***Reggio nell'Emilia***	**Venezia**
Firenze	**Cagliari**	***Terni***	**Trento**
Sassari	**Livorno**	***Bolzano***	**Brescia**
Verona	**Pescara**	***Padova***	**Padova**
Prato	**Padova**	**Perugia**	**Ravenna**
Trieste	**Reggio nell'Emilia**	**Salerno**	**Trieste**
Genova	**Firenze**	**Bologna**	**Sassari**
Piacenza	**Ravenna**	**Verona**	**Firenze**
Ravenna	**Bolzano**	**Piacenza**	**Terni**
Taranto	**Perugia**	**Modena**	**Pescara**
Monza	**Trento**	**Sassari**	**Monza**
Vicenza	**Palermo**	**Livorno**	**Parma**
Ancona	**Prato**	**Ancona**	**Roma**
Modena	**Piacenza**	**Cagliari**	**Livorno**
Reggio di Calabria	**Vicenza**	**Roma**	**Bari**
Livorno	**Sassari**	**Ravenna**	**Perugia**
Ferrara	**Reggio di Calabria**	**Pescara**	**Genova**
Brescia	**Taranto**	**Trieste**	*Cagliari*
Reggio nell'Emilia	**Bergamo**	**Bari**	*Ancona*
Bergamo	*Verona*	**Forlì**	*Verona*
Forlì	*Ancona*	*Genova*	*Salerno*
Terni	*Foggia*	*Reggio di Calabria*	*Palermo*
Padova	*Trieste*	*Palermo*	*Reggio di Calabria*
Parma	*Genova*	*Taranto*	*Catania*
Trento	*Catania*	*Catania*	*Foggia*
Perugia	*Brescia*	*Foggia*	*Taranto*

Development Goals (Heads of State and Government and High Representatives 2015). In this paper, the UCI model is used to investigate waste management in the 40 largest Italian municipalities showing its main characteristics and its potential to represent multidimensional hierarchical phenomena. Therefore, the model provides a hierarchical system of CIs and corresponding rankings, which might be used for policy actions. An additional analysis that excludes the effect of two important external variables, namely *Density* and *Touristic rate*, shows another important feature of the model.

Appendix A: Estimation of the parameters of $\mathbf{R}_{\mathrm{EU}}$

The estimates of $\mathbf{R}_{\mathrm{W}}$, $\mathbf{R}_{\mathrm{B}}$ and $\mathbf{V}$ provided in the following are obtained by minimizing Eq. (8) subject to constraints (3)–(7).

(a) Estimation of $\mathbf{R}_{\mathrm{W}}$: for fixed $\widehat{\mathbf{V}}$,

$$\widehat{\mathbf{R}}_{\mathrm{W}} = \mathrm{diag}\left(\widehat{\mathbf{V}}'(\mathbf{R} - \mathbf{I}_p)\widehat{\mathbf{V}}\right)\left((\widehat{\mathbf{V}}'\widehat{\mathbf{V}})^2 - \widehat{\mathbf{V}}'\widehat{\mathbf{V}}\right)^{-1}.$$

$\widehat{\mathbf{R}}_{\mathrm{W}}$ minimizes Eq. (8), given $\widehat{\mathbf{R}}_{\mathrm{B}}$ and $\widehat{\mathbf{V}}$, and satisfies condition (5). It should be noted that since the diagonal of $\widehat{\mathbf{R}}_{\mathrm{B}}$ is set to zero by constraint (6), it does not affect the estimates of $\mathbf{R}_{\mathrm{W}}$. The inverse of $(\widehat{\mathbf{V}}'\widehat{\mathbf{V}})^2 - \widehat{\mathbf{V}}'\widehat{\mathbf{V}}$ results from the fact that $\widehat{\mathbf{V}}'\widehat{\mathbf{V}}$ is a diagonal matrix whose diagonal entries represent the group sizes, and the Moore-Penrose inverse of a matrix $\mathbf{M}$, that is, $\mathbf{M}^{+}$, is equal to the inverse of the same matrix, that is, $\mathbf{M}^{-1}$, if $\mathbf{M}$ is diagonal.

(b) Estimation of $\mathbf{R}_{\mathrm{B}}$: for fixed $\widehat{\mathbf{V}}$, $\widehat{\mathbf{R}}_{\mathrm{B}}$ is calculated as the closest matrix to

$$\tilde{\mathbf{R}}_{\mathrm{B}} = \widehat{\mathbf{V}}^{+}\mathbf{R}(\widehat{\mathbf{V}}^{+})'$$

in the LS sense that satisfies condition (6). Indeed, the off-diagonal elements of $\tilde{\mathbf{R}}_{\mathrm{B}}$ simply denote the correlations between Q variable groups, but they do not necessarily satisfy the ultrametric condition. An average linkage (UPGMA, Sokal and Michener 1958) algorithm for correlations can be used to compute $\widehat{\mathbf{R}}_{\mathrm{B}}$.

(c) Estimation of $\mathbf{V}$: for fixed $\widehat{\mathbf{R}}_{\mathrm{W}}$ and $\widehat{\mathbf{R}}_{\mathrm{B}}$, each row of $\mathbf{V}$, that is, $\mathbf{v}_j$, $j = 1, \dots, p$, is estimated by fixing the remaining rows and setting

$$\begin{cases} \hat{v}_{jq} = 1 & \text{if} \quad q = \underset{q'=1,\dots,Q}{\arg\min} F(\widehat{\mathbf{R}}_{\mathrm{W}}, \widehat{\mathbf{R}}_{\mathrm{B}}, [\hat{\mathbf{v}}_1, \dots, \mathbf{v}_j = \mathbf{i}_{q'}, \dots, \hat{\mathbf{v}}_p]') \\ \hat{v}_{jq} = 0 & \text{otherwise} \end{cases}$$

where $\mathbf{i}_q$ is the qth row of the identity matrix of order Q. Therefore, estimating the rows of $\mathbf{V}$ corresponds to assigning each variable to only one of the Q disjoint groups (conditions 3 and 4) to minimize the loss function.

Appendix B: The UCI model algorithm

The algorithm for the estimation of the UCI model is provided in Algorithm 1. The code for Algorithm 1 is written in MATLAB and is available upon request to the authors.

Algorithm 1: The UCI model algorithm

```
Input: $\mathbf{X}$, $Q$
Fixed values $\epsilon$: convergence tolerance value (default: $0.1^6$);
maxiter: maximum number of iterations (default: $100$);
$\alpha$: test significance level (default: $0.05$);
rndstart: number of random starts (default: $100$);
constrV: membership matrix of dimension $(p \times Q)$ for semi-confirmatory analysis
  (default: null matrix for exploratory analysis);
for $i = 1$ to rndstart do
    Initialization $\widehat{\mathbf{V}}^{(0)}$ = constrV and corresponding estimates $\widehat{\mathbf{R}}_{\mathrm{W}}^{(0)}$ and $\widehat{\mathbf{R}}_{\mathrm{B}}^{(0)}$. If
      constrV is null, $\widehat{\mathbf{V}}^{(0)}$ is a random initial partition;
    if Constraint (7) does not hold then
        $min\{{}_{W}r_{qq}^{(0)} : q = 1, \ldots, Q\} \leftarrow max\{{}_{B}r_{qh}^{(0)} : q, h = 1, \ldots, Q, h \neq q\}$;
    end
    while $F_{diff}^{(t)} > \epsilon$ and $t \leq$ maxiter do
        Step 1 Update $\widehat{\mathbf{V}}$
            $\widehat{\mathbf{V}}^{(t)}$ is computed according to point (c) in Appendix A.
        endStep1
        Step 2 Update $\widehat{\mathbf{R}}_{\mathrm{W}}$
            $\widehat{\mathbf{R}}_{\mathrm{W}}^{(t)}$ is computed according to point (a) in Appendix A and the
              configuration of $\widehat{\mathbf{V}}^{(t)}$.
        endStep2
        Step 3 Update $\widehat{\mathbf{R}}_{\mathrm{B}}$
            $\widehat{\mathbf{R}}_{\mathrm{B}}^{(t)}$ is computed according to point (b) in Appendix A and the
              configuration of $\widehat{\mathbf{V}}^{(t)}$.
        endStep3
        $F^{(t)} \leftarrow F(\widehat{\mathbf{R}}_{\mathrm{W}}^{(t)}, \widehat{\mathbf{R}}_{\mathrm{B}}^{(t)}, \widehat{\mathbf{V}}^{(t)})$ in Eq. (8);
        $F_{diff}^{(t)} \leftarrow (F^{(t-1)} - F^{(t)})$;
    end
end
Test The significance between two sequential values of $\widehat{\mathbf{R}}_{\mathrm{B}}$ (from the lowest value
  upwards) is assessed according to the test reported in Section 3.2;
Computation of the SCIs and GCI The first-level, higher-level SCIs and the
  GCI are constructed as reported in Section 3.3;
Output: The hierarchy with first-level SCIs, higher-order SCIs corresponding to
        statistically significant levels and the GCI.
```

Supplementary Information The online version of this article (https://doi.org/10.1007/s00180-023-01333-9) contains supplementary material, which is available to authorized users.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. The authors did not received support from any organization for the submitted work.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Anderson TW, Rubin H (1956) Statistical inferences in factor analysis. Proceedings of the Third Symposium on Mathematical Statistics and Probability 5:111–150

Cailliez F (1983) The analytical solution of the additive constant problem. Psychometrika 48(2):305–308

Cattell RB (1978) Higher-order factors: models andormulas. Springer, US, Boston, MA, pp 192–228

Cavicchia C, Vichi M (2021) Statistical model-based composite indicators for tracking coherent policy conclusions. Soc Indic Res 156(2):449–479

Cavicchia C, Vichi M (2022) Second-order disjoint factor analysis. Psychometrika 87(1):289–309

Cavicchia C, Vichi M, Zaccaria G (2020) The ultrametric correlation matrix for modelling hierarchical latent concepts. Adv Data Anal Classif 14(4):837–853

Cavicchia C, Sarnacchiaro P, Vichi M (2021) A composite indicator for the waste management in the eu via hierarchical disjoint non-negative factor analysis. Socio-Econ Plan Sci 73:100832

Cavicchia C, Vichi M, Zaccaria G (2022) Gaussian mixture model with an extended ultrametric covariance structure. Adv Data Anal Classif 16(2):399–427

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334

Dellacherie C, Martinez S, San Martin J (2014) Inverse M-matrices and ultrametric matrices. Lecture Notes in Mathematics, Springer International Publishing

Diaz-Farina E, Díaz-Hernández JJ, Padrón-Fumero N (2020) The contribution of tourism to municipal solid waste generation: A mixed demand-supply approach on the island of Tenerife. Waste Manage 102:587–597

Dunn J, Clark VA (1969) Correlation coefficients measured on the same individuals. J Am Stat Assoc 64(325):366–377

European Commission (2010) Commission regulation (EU) No 849/2010 of 27 September 2010 amending Regulation (EC) No 2150/2002 of the European parliament and of the Council on waste statistics. Off J Eur Union 53(L 253):2–41, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32010R0849 &from=EN

European Parliament, Council of the European Union (1999) Council Directive 1999/31/EC of 26 April 1999 on the landfill of waste. Off J Eur Communities 42(L 182):1–39, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:1999:182:FULL &from=EN

European Parliament, Council of the European Union (2018) Directive (EU) 2018/850 of the European Parliament and the Council of 30 May 2018 amending Directive 1999/31/EC on the landfill of waste. Off J Eur Union 61(L 150):100–108, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2018:150:TOC

Fisher RA (1921) On the probable error of a coefficient of correlation deduced from a small sample. Metron 1:3–32

Heads of State and Government and High Representatives (2015) Transforming our world: the 2030 agenda for sustainable development, a/res/70/1. Tech. rep., United Nations, https://sustainabledeve lopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24(6):417–441, 498–520

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218

Hunter MA, Takane Y (2002) Constrained principal component analysis: various application. J Educ Behav Stat 27(2):105–145

Kaiser HF (1960) The application of electronic computers to factor analysis. Educ Psychol Meas 20(1):141–151

Kline P (2000) The handbook of psychological testing, 2nd edn. Routledge

Matai K (2015) Sustainable tourism: waste management issues. J Bas Appl Eng 2(1):1445–1448

Mateu-Sbert J, Ricci-Cabello I, Villalonga-Olives E, Cabeza-Irigoyen E (2013) The impact of tourism on municipal solid waste generation: The case of Menorca Island (Spain). Waste Manage 33(12):2589–2593

Nardo M, Saisana M, Saltelli A, Tarantola S (2005) Tools for composite indicators building. Tech. Rep. EUR 21682, Join Research Centre, Ispra, Italy, https://knowledge4policy.ec.europa.eu/publication/tools-composite-indicators-building-0_en

OECD-JRC (2008) Handbook on constructing composite indicators: Methodology and user guide. Tech. rep., OECD Publishing, https://www.oecd.org/sdd/42495745.pdf

Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine and Journal of Science 2(11):559–572

Schmid J, Leiman JM (1957) The development of hierarchical factorial solutions. Psychometrika 22(1):53–61

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 38(2):1409–1438

Steiger JH (1980) Tests for comparing elements of a correlation matrix. Psychol Bull 87(2):245–251

Takane Y, Shibayama T (1991) Principal component analysis with external information on both subjects and variables. Psychometrika 56(1):97–120

Wherry RJ (1959) Hierarchical factor solutions without rotation. Psychometrika 24(1):45–51