Socio-Economic Planning Sciences 105 (2026) 102514

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

journal homepage: www.elsevier.com/locate/seps

Robust interactive and interpretable hierarchical PCA: Applications to crime data[☆]

Marco Tolla [a], Marco Riani [b,*], Paolo Mariani [c], Corrado Crocetta [d]

[a] Guardia di Finanza (Italian Economic and Financial Police), Viale XXI Aprile, 51, Rome, 00162, Italy
[b] Department of Economics and Management, University of Parma, Via Kennedy 6, Parma, 43123, Italy
[c] Dipartimento di Economia, Metodi Quantitativi e Strategie di Impresa, Universita' degli Studi di Milano Bicocca, Via Bicocca degli Arcimboldi, 8, Milan, 20126, Italy
[d] Dipartimento di Ricerca e innovazione umanistica, Universita' degli Studi di Bari Aldo Moro, Str. della Torretta, 70122, Bari, 20126, Italy

ARTICLE INFO

Keywords:
Composite indicators
Robust methods
Dimension reduction
Interpretability
Interactivity

ABSTRACT

In a context characterized by increasing institutional complexity and growing constraints on available resources, this paper proposes an interpretable framework for the construction of composite indicators aimed at supporting strategic decision making in public administration with specific emphasis on financial crime data. Robust statistical methods provide valuable support for analysing how and where resources should be deployed, enabling data-driven decisions aimed at improving the performance of Public Administrations for the benefit of the community. However, such analyses must produce results that are interpretable, clearly quantify the contribution of each unit, and remain robust in the presence of atypical observations. Furthermore, the inclusion of interactive tools is essential to allow end users to easily explore the characteristics and relative positions of units within the analytical space. Motivated by these requirements, this paper describes the interactive and interpretable framework based on robust hierarchical principal component analysis for optimizing resources and operational strategies within the Guardia di Finanza (GdF - Italian Economic and Financial Police).

1. Introduction

Since the early 1930s, statistical methods have been systematically employed to study crime in terms of its magnitude, dynamics, determinants, treatment, and prevention [1]. In their early applications, however, crime and social statistics were characterized by heterogeneous methodological standards and limited analytical consistency. Already more than fifty years ago, the construction of one of the earliest and most influential national crime indices raised critical issues concerning the definition, selection, and weighting of its constituent components, calling for internal methodological revisions [2].

A recurring challenge in the analysis of official crime statistics lies in the aggregation of heterogeneous offence categories into a single synthetic measure. Common strategies include the construction of separate indices for offences against persons and offences against property, or the aggregation of offence categories after weighting them according to their perceived severity. In this context, Parker and McDowall [3] proposed the use of factor analysis techniques to derive composite crime indices. More broadly, the measurement of crime severity has been extensively debated, with particular attention devoted to the methodological requirements underlying magnitude estimation approaches [4].

Over time, the main sources of official crime statistics in the United Kingdom, Europe, and the United States have been subject to persistent concerns regarding their reliability for both descriptive and predictive purposes. These limitations primarily arise from under-reporting by victims and, more generally, from structural data quality issues [5]. More recently, the widespread diffusion of digital technologies has enabled the collection and storage of increasingly large volumes of spatio-temporal crime data. This development has created new opportunities for the application of advanced statistical and data-driven methods aimed at identifying spatial and temporal crime patterns, extracting predictive models of criminal activity, and supporting the design of more effective crime prevention strategies [6].

When dealing with large, high-dimensional, and heterogeneous data structures, an adequate level of synthesis becomes essential. The joint analysis of multiple dimensions characterizing complex phenomena can be effectively addressed through statistical methodologies leading to

☆ This article is part of a Special issue entitled: 'Statistical Measurement of Economic Analysis' published in Socio-Economic Planning Sciences.

* Corresponding author.
E-mail addresses: mriani@unipr.it (M. Riani), paolo.mariani@unimib.it (P. Mariani), corrado.crocetta@uniba.it (C. Crocetta).

https://doi.org/10.1016/j.seps.2026.102514
Received 20 March 2026; Received in revised form 14 May 2026; Accepted 19 May 2026
Available online 25 May 2026

the construction of composite indicators [7,8]. With specific reference to Italy, several summary indices related to crime and socio-economic conditions are published on a regular basis. A notable example is the annual provincial-level quality-of-life assessment published by the Italian economic newspaper *Il Sole 24 Ore*. Within this framework, Gismondi and Russo [9] proposed a methodological approach based on variable pre-processing, redundancy analysis, and structured weighting schemes.

Beyond their descriptive role, such composite indicators also represent a crucial analytical support for public institutions that operate in complex and territorially heterogeneous environments and are required to allocate resources and define priorities in a transparent and evidence-based manner. In this perspective, the construction of synthetic, interpretable measures becomes particularly relevant for law enforcement and economic police bodies.

The *Guardia di Finanza* (GdF) is a modern military police force entrusted by law with a broad and highly articulated set of functions, that forms an integral part of the Italian Armed Forces and exercises general competence in economic and financial matters. Its institutional mandate includes, among other activities, the prevention, investigation, and repression of financial and economic violations; maritime surveillance for police purposes; assistance and reporting in the maritime domain; monitoring compliance with provisions of political and economic relevance; contribution to the military defense of the State and the maintenance of public order and security; safeguard of the financial interests of the State, Italian Regions, local authorities, and the European Union; fight against counterfeiting of currencies, securities, and means of payment; monitoring of financial flows and capital movements; protection of financial and securities markets, credit systems, and intellectual and industrial property rights; prevention and repression of organized crime, money laundering, and terrorist financing. Moreover, through its naval and air components, the GdF plays a crucial role in combating illegal trafficking by sea.

The breadth, relevance, and sensitivity of these functions require effective analytical tools to support the optimal allocation of resources and the definition of sound operational strategies. In this context, statistical methods play a fundamental role in analysing the pronounced territorial heterogeneity characterizing the Italian context. To this end, a comprehensive system of indicators has been developed to describe and compare institutional, social, economic, and criminal dimensions at the provincial level, covering the Italian provinces.

The aim of this paper is to illustrate the procedure adopted by the GdF to construct an Index of Socio-Economic and Criminal Complexity and Relevance that is robust, interpretable, and methodologically sound, while remaining easily understandable for end users. Particular emphasis is placed on interpretability and transparency, explicitly avoiding the use of black-box algorithms.

The remainder of the paper is organized as follows. Section 2 links the proposed approach to the literature on hierarchical principal component analysis and composite indicators. Section 3 provides a concise description of the dataset, which combines internal administrative data with external sources. Section 4 details the aggregation procedures adopted to compute the final composite indicator. Section 5 describes the interactive robust methodology employed at each aggregation stage. Finally, Section 6 concludes the paper.

2. Hierarchical robust PCA within the landscape of composite indicator methodologies

The construction of composite indicators through Principal Component Analysis (PCA) is well established in the social sciences, where it is routinely employed to synthesize large sets of correlated variables into a limited number of latent dimensions [10,11]. In applications concerning socio-economic development, institutional quality, territorial disparities and crime analysis, PCA offers a data-driven alternative to purely normative weighting schemes, allowing the data structure itself to determine the relative importance of indicators. In this sense, PCA-based aggregation enhances statistical objectivity while preserving dimensional parsimony.

However, when indicators span heterogeneous conceptual domains, a single global PCA often produces components that are difficult to interpret and potentially unstable in the presence of noise. To address these limitations, hierarchical or multi-stage PCA strategies have been proposed [12]. In such approaches, variables are first organized into substantively coherent domains and reduced within each block; subsequently, intermediate components are further aggregated through additional PCA steps. This structured procedure preserves the variance-extraction properties of PCA while maintaining interpretability and traceability across aggregation levels. Hierarchical PCA has proven particularly suitable in regional analysis, well-being measurement, environmental performance assessment, and institutional evaluation, contexts in which transparency of the aggregation process is essential. Moreover, this structured strategy aligns with the methodological guidelines for composite indicator construction promoted by international organizations such as the OECD and the European Commission [13], which emphasize careful domain definition, explicit aggregation schemes, and sensitivity analysis with respect to weighting and normalization choices.

Beyond its statistical foundations, the hierarchical robust PCA framework adopted here is grounded in a broader epistemological perspective on the measurement of multidimensional phenomena. From the viewpoint of Complexity Theory, as articulated by Nobel laureate Giorgio Parisi [14,15], complex systems arise from the interaction of numerous heterogeneous components whose collective behaviour cannot be reduced to the sum of their parts. Parisi's work on spin glasses and collective dynamics illustrates how simple local interactions generate emergent macro-level patterns. This insight is directly applicable to socio-economic and crime systems, where territorial outcomes emerge from the interplay of economic conditions, institutional structures, and law-enforcement activities. The hierarchical architecture operationalises this perspective by preserving the multi-level structure of the phenomenon while extracting dominant latent dimensions, thus respecting the emergent nature of socio-economic and criminal complexity.

A well-known limitation of classical PCA in applied socio-economic and administrative contexts is its sensitivity to atypical observations and leverage points [16]. Crime and law-enforcement data often display heavy-tailed distributions, structural breaks, and local anomalies that may distort covariance structures and consequently bias loadings and scores. Robust extensions of PCA mitigate this problem by relying on high-breakdown covariance estimators or projection-pursuit techniques, thereby limiting the influence of outliers while preserving the dominant latent structure. Embedding robust PCA within a hierarchical aggregation scheme ensures stability at each stage of dimensionality reduction and enhances the coherence of the overall indicator system.

An additional methodological innovation concerns the introduction of explicit sign constraints on loadings. By imposing non-negativity restrictions, the latent dimension is forced to reflect a coherent substantive interpretation, consistent with the assumption that all selected variables contribute positively to the underlying construct of socio-economic and criminal complexity. This feature strengthens interpretability and distinguishes the proposed framework from standard hierarchical PCA implementations.

Within the broader landscape of composite indicator methodologies, the proposed hierarchical robust PCA occupies a distinct position. Partial Least Squares Path Modelling (PLS-PM), for instance, has been extensively developed for estimating models containing composites of composites through repeated indicators and two-stage approaches [17,18]. Applications such as Crocetta et al. [19] demonstrate its flexibility in handling higher-order latent constructs with formative and reflective specifications. However, PLS-PM relies on explicitly specified structural paths and variance-based structural equation modelling, whereas the present approach remains entirely data-driven and does not impose *a priori* causal assumptions. This makes hierarchical robust PCA particularly suitable for exploratory contexts where theoretical path structures are unavailable or contested.

The Mazziotta–Pareto Index (MPI) family represents a different paradigm, grounded in non-compensatory aggregation principles [20]. Unlike PCA, which allows compensation across dimensions, MPI penalizes unbalanced profiles through dispersion-adjusted aggregation rules. While MPI emphasizes equity and interpretability, hierarchical robust PCA prioritizes variance extraction and statistical optimality, with sign constraints ensuring directional coherence. These approaches reflect alternative normative positions in the compensation–penalisation trade-off.

Partially Ordered Set (POSET) theory offers yet another methodological alternative [21]. POSET methods avoid weighting and aggregation functions altogether, deriving rankings from dominance relations implied by indicator profiles. Although this guarantees maximal transparency and minimal arbitrariness, it yields ordinal information only and becomes computationally demanding with many units. By contrast, hierarchical robust PCA provides cardinal composite scores with explicit linear coefficients, facilitating policy decisions that require interval-level comparability.

The contribution of this paper lies in synthesizing these perspectives. The proposed framework preserves the variance-exploiting properties of PCA, incorporates robustness against outliers, ensures interpretability through sign constraints, and embeds the entire procedure within an interactive graphical environment designed for policy use. In the taxonomy of composite indicator methodologies, it can thus be viewed as more structured than global PCA, more flexible than MPI, more scalable than POSET, and more transparent than PLS-PM, offering a coherent strategy for measuring complex territorial phenomena.

3. Data description

The spatial unit of analysis adopted in this study is the Italian administrative province. A total of 106 territories are considered, corresponding to the NUTS 3 level [22]. The number of provinces is 106 because the province of South Sardinia has been included within the province of Cagliari [23] for the purposes of the current work. The indices are defined with reference to the most recent available period. Since not all variables are observed over the same time span, missing temporal alignments were addressed by estimating trends in order to harmonize the reference period across all indicators.

The resulting data structure can therefore be represented as a two-way array defined by observations (106 provinces), variables (319 indicators). In future developments, this structure will be extended to include a temporal dimension. The analytical framework adopted is that of principal component analysis for multi-table and multi-block data. In particular, the proposed methodology largely follows the Multiple Factor Analysis (MFA) approach [12]. MFA is specifically designed for situations in which different sets of variables are measured on the same set of observations. While the number and nature of variables may vary across groups, the observational units must remain identical.

In the present application, the different data sets consist of measurements collected on the same observational units (Italian provinces) across different sources and reference years. The core idea underlying MFA is to normalize each data block through an initial principal component analysis, so that the first principal components of each block can be combined into a common representation of the observations, often referred to as a compromise or consensus space [24]. The objective is thus to integrate heterogeneous groups of variables describing the same phenomenon.

A preliminary step of the analysis consists in verifying whether the selected variables adequately represent distinct underlying phenomena. Data quality was ensured through direct acquisition from the original sources, following a methodological approach inspired by the Quality Information System developed by the Italian National Institute of Statistics (ISTAT) [25]. Detailed metadata were maintained to document survey methodologies, data processing procedures, and quality assurance activities. Microdata integration was carried out using exact matching procedures across the different data archives [26]. Whenever possible, internal consistency checks were performed by comparing information drawn from different sources or by analysing historical time series of the same variables. Particular attention was devoted to derived variables, especially those involving ratios or relationships between indicators, for which standardization procedures were applied when necessary.

The production cycle of the composite indices is periodic and reflects both operational constraints and evolving informational needs. Since many of the variables are strongly correlated with dimensional factors such as population size and the number of firms, two alternative analytical strategies were considered. The first involved standardizing variables with respect to these dimensional factors, while the second retained the original scales in order to explicitly preserve dimensional effects. In this paper, we focus on the results obtained from the latter approach, as it allows us to capture the overall magnitude of the phenomena under study and to reflect the structural differences across provinces. Moreover, preserving dimensional effects is particularly relevant for policy-oriented interpretation, where absolute levels, rather than normalized measures, are often of primary interest.

The project began with the construction of a large, reliable, and systematically organized dataset. Through collaboration with governmental, economic, financial, and other institutions responsible for the census, collection, and archiving of statistical information, an automated system for the acquisition of data for each reference year was established. This system was further enriched with statistical information collected directly by the *Guardia di Finanza* in the course of its institutional activities.

The primary objective was to define automated procedures capable of recording and certifying the origin, nature, content, acquisition methods, and timing of each individual datum used in the statistical analyses, ensuring that:

- the information employed can always be verified over time;
- the data can be periodically updated, allowing for the construction of stable, coherent, and complete historical series.

At present, more than one thousand variables are collected. These variables are highly heterogeneous in terms of nature, measurement units, and order of magnitude, and each contributes to highlighting specific institutional, economic, social, or criminal dimensions deemed relevant. For the construction of the indicators, 715 variables were selected based on their significance and were grouped into two main sets:

- the *external context*, consisting of data produced by other institutions;
- the *internal context*, comprising statistical information directly collected by the *Guardia di Finanza*.

Variables in the external context mainly describe institutional, economic, and social features that are generally independent of social deviance phenomena. Conversely, variables in the internal context primarily reflect manifestations of social deviance linked to economic–financial crime and organized crime, as they derive from statistics on police activity.

This distinction allows assessment of how the perceived relevance of territorial realities changes when considering:

- only the statistical data available from other institutions (319 external-context variables);
- only the information collected by the *Guardia di Finanza* (396 internal-context variables);
- the combined use of both information sources.

4. Construction of the index of socio-economic and criminal complexity and relevance

Given the GDF's tasks, it is important to be able to compare different territories from various, even highly heterogeneous, perspectives. Thus, it was necessary to consider the relevance and complexity of each province at the institutional level (the various public bodies present in the territory and the related public functions exercised there); at the economic level (the different economic activities conducted and the wealth produced); at the social level (the specific non-economic characteristics of the different territorial realities); and, finally, at the level of the criminal phenomena detected. The set of all these characteristics can be defined, for the purposes of this study, as the socio-economic and criminal relevance and complexity of the territory considered, determined by the comprehensive analysis of all 715 statistical variables used.

Let $n = 106$ denote the number of territorial reference units (Italian provinces), and let $\mathbf{X}^{(E)} = \{X_j^{(E)}\}_{j=1}^{319}$ and $\mathbf{X}^{(I)} = \{X_j^{(I)}\}_{j=1}^{396}$ denote, respectively, the matrices of external and internal variables observed for each unit. External variables describe institutional, social, and economic characteristics (232) and crime related components (87) derived from administrative and official statistical sources. Internal variables are collected directly by the *Guardia di Finanza* and provide a unique observational perspective on economic, financial, and organized crime phenomena.

The construction of the *Index of Socio-Economic and Criminal Complexity and Relevance* (ISCCER) follows a hierarchical aggregation scheme. At each stage of the procedure, dimensionality reduction is performed using *constrained robust principal component analysis* (RCPCA), in order to limit the influence of atypical observations and ensure the stability and interpretability of the resulting indicators.

Socio-economic component (external data). The 232 external variables describing institutional, social, and economic dimensions unrelated to social deviation are first grouped into $K_1 = 24$ homogeneous domains, each representing a specific area of analysis (e.g. socio-demographic characteristics, judicial and administrative institutions, education and employment, income, household consumption, vehicle ownership, public expenditure and procurement). Let $\mathbf{X}_k$ denote the standardized data matrix associated with the kth domain ($k = 1, \dots, 24$). For each domain, a sectoral composite indicator is obtained as the first robust principal component:

$$Z_k^{(E)} = \mathbf{X}_k \mathbf{v}_k^{(1)},$$

where $\mathbf{v}_k^{(1)}$ is the eigenvector associated with the largest robust eigenvalue of the corresponding correlation matrix. The 24 domain indicators are then aggregated by applying robust PCA to the standardized matrix $\mathbf{Z}^{(E)} = (Z_1^{(E)}, \dots, Z_{24}^{(E)})$, yielding the *Index of Socio-Economic Complexity* (ISC).

Crime-related component (external data). The remaining 87 external variables, describing social deviation phenomena and critical conditions in the labour market, credit system, and business sector, are grouped into $K_2 = 10$ crime-related domains (e.g. organized and economic–financial crime, crimes against public integrity, robberies, thefts, business crises, labour market vulnerabilities). For each domain $j = 1, \dots, 10$, a sectoral indicator is obtained via robust PCA:

$$Z_j^{(C)} = \mathbf{Y}_j \mathbf{u}_j^{(1)},$$

where $\mathbf{Y}_j$ is the standardized data matrix for the jth domain. The resulting indicators are then aggregated using robust PCA to obtain the *Index of Economic and Financial Crime* (ICF).

Overall index (external data). The two synthetic indicators ISC and ICF are finally combined by applying robust PCA to their standardized matrix,

$$\mathbf{Z}^{(EC)} = (\mathrm{ISC}, \mathrm{ICF}),$$

and retaining the first component, which defines the ISCCER based on external information. Fig. 1 graphically summarizes the hierarchical construction of ISCCER.

Extension to internal and integrated data. The same hierarchical procedure is applied to the internal dataset $\mathbf{X}^{(I)}$, producing a system of sectoral and composite indicators describing territorial complexity solely on the basis of law-enforcement information. Subsequently, the procedure is applied to the integrated dataset

$$\mathbf{X} = (\mathbf{X}^{(E)}, \mathbf{X}^{(I)}),$$

comprising all 715 variables. In this integrated setting, 47 additional sectoral indicators are constructed to capture specific forms of social, economic, and financial deviance observable only through internal operational data.

Weighted robust aggregation. At higher aggregation levels, robust PCA can be applied either to unweighted sectoral indicators or to indicators weighted according to their statistical relevance. Let Z_k denote a generic sectoral indicator and let λ_k be the proportion of robust variance explained by its first component. A variance-based weighting scheme is defined as

$$w_k = \frac{\lambda_k}{\sum_{h=1}^{K} \lambda_h},$$

while a size-adjusted scheme accounting for the number p_k of original variables in each cluster is given by

$$w_k^* = \frac{\lambda_k p_k}{\sum_{h=1}^{K} \lambda_h p_h}.$$

Robust PCA is then applied to the weighted indicators $w_k Z_k$ or $w_k^* Z_k$ to obtain higher-level composite indices.

The availability of three distinct indicator systems based on external data only, internal data only, and on the integrated dataset enables a systematic comparison of territorial rankings. This comparison serves two complementary objectives. First, it allows the assessment of the robustness and stability of provincial rankings with respect to the underlying information base. Second, it permits the identification of those provinces whose relative positions change most markedly when internal data are incorporated, thereby highlighting territorial contexts whose socio-economic and criminal complexity is not fully captured by external indicators alone. Overall, the three indicator systems exhibit substantial concordance, suggesting that the main structural features of territorial complexity are consistently captured across different information bases.

Moreover, the proposed framework allows for the adoption of alternative weighting schemes in the aggregation stages. In particular, weights may account for the proportion of variance explained by each sectoral indicator and/or for the number of original variables contributing to each domain. This flexibility enables the *Guardia di Finanza* to conduct an extensive sensitivity analysis and to evaluate the stability of the results under different plausible aggregation assumptions.

A further relevant aspect concerns the interpretability of the final composite indicators. At each aggregation step, robust PCA produces loading vectors that, possibly combined with the selected weighting scheme, determine the contribution of lower level indicators to higher level indices. As a consequence, the final composite indicator obtained from the integrated dataset can be expressed as a linear combination of all original standardized variables. Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$ denote the vector of standardized variables for province i, where p is the total

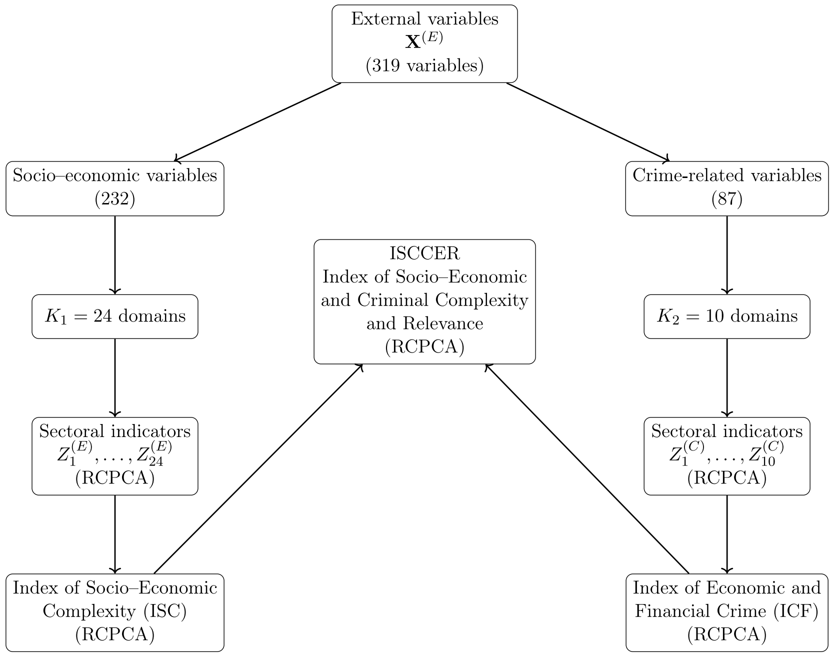

Fig. 1. Hierarchical construction of the Index of Socio-Economic and Criminal Complexity and Relevance (ISCCER) based on external data. At each aggregation stage, dimensionality reduction is performed using constrained robust principal component analysis (RCPCA).

number of variables considered. The final index for province i can then be written as

$$\mathrm{ISCCER}_i = \sum_{j=1}^{p} \alpha_j x_{ij} = \boldsymbol{\alpha}^\top \mathbf{x}_i,$$

where the coefficients α_j result from the product of the loading vectors obtained at each aggregation level and incorporate the chosen weighting scheme.

This representation makes it possible to explicitly assess the contribution of each original variable to the overall index. Analogously, the contribution of each territorial unit to the determination of the final aggregation structure can be evaluated. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$ denote the vector of global coefficients. The influence of province i on the final index can be summarized by

$$C_i = \boldsymbol{\alpha}^\top \mathbf{x}_i, \tag{1}$$

which coincides with the province-specific score on the first robust principal component and allows analysts to understand how individual territorial profiles shape the overall distribution of the composite indicator.

In the proposed framework, the entire set of $p = 715$ standardized variables is selected so as to exhibit a *non-negative relation* with the latent dimension of socio-economic and criminal complexity captured by the composite indicator. Consequently, negative loadings arising at any aggregation stage are not interpreted as evidence of an inverse substantive relationship, but rather as artefacts induced by atypical observations or leverage points or by problems in the data collection procedure, which may affect the orientation of the first principal component.

Formally, let $\mathbf{v}^{(1)} = (v_1^{(1)}, \ldots, v_p^{(1)})^\top$ denote the loading vector associated with the first (robust) principal component obtained at a generic aggregation step. A sign constraint is imposed by defining a modified loading vector $\tilde{\mathbf{v}}^{(1)}$ such that

$$\tilde{v}_j^{(1)} = \begin{cases} v_j^{(1)}, & \text{if } v_j^{(1)} \geq 0, \\ 0, & \text{if } v_j^{(1)} < 0, \end{cases} \qquad j = 1, \ldots, p. \tag{2}$$

This constraint is enforced systematically at all levels of the hierarchical aggregation procedure.

As a consequence, in the final linear representation of the composite indicator in (1), the coefficient vector $\boldsymbol{\alpha}$ satisfies

$$\alpha_j \geq 0 \quad \text{for all } j = 1, \ldots, p,$$

with $\alpha_j = 0$ whenever the corresponding loading is negative at any aggregation stage. Hence, variables whose sign is not consistent with the assumed direction of association with criminal complexity do not contribute to the index. As noted by one of the referees, this constraint breaks the orthogonality with respect to the remaining latent dimensions. While this observation is correct, it does not represent a limitation in the present framework, since at each step of the hierarchical procedure only the first principal component is retained.

This formulation yields several analytical advantages. First, it induces a natural ranking of all $p = 715$ variables according to their global contribution α_j to the composite indicator. Second, it allows the explicit identification of variables whose contribution is null, thereby highlighting dimensions that do not affect the latent index of criminality and socio-economic complexity. Third, it enables an *ex post* quantitative assessment of the relative importance of different thematic domains, as well as of the contribution of internal versus external indicators to the overall index, through suitable aggregation of the coefficients α_j at the sectoral level.

Finally, the province-specific score (1) provides a direct measure of the contribution of each territorial unit to the composite indicator. This representation makes it possible to evaluate, for each province, the marginal impact of including internal indicators in the aggregation process. In particular, it allows the identification of provinces for which internal data substantially alter the composite score, as well as those for which the inclusion of internal indicators produces negligible changes, thereby offering valuable insights for targeted operational and strategic decision making.

Flexible indicator construction. Beyond the predefined indicator system, subsets of variables have been recombined to construct ad hoc composite indicators tailored to specific analytical or operational objectives. Robust PCA is consistently applied at each aggregation stage.

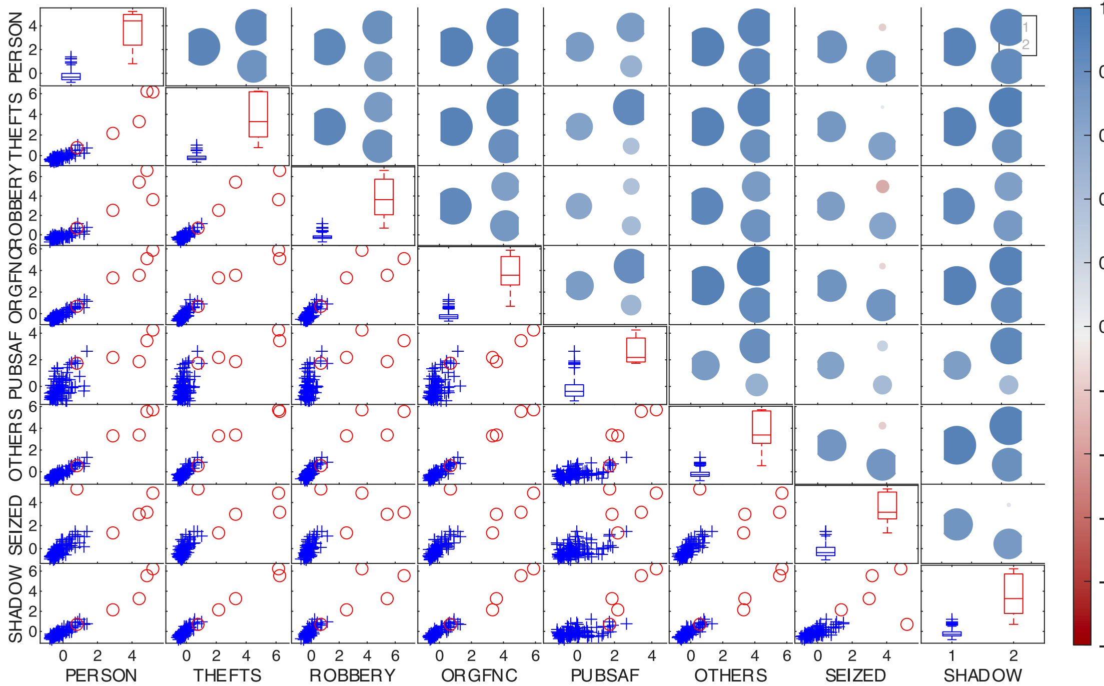

Fig. 2. Scatterplot matrix with detected multivariate outliers. Univariate boxplots are shown on the diagonal, while the lower triangular panels display bivariate scatterplots with outliers highlighted by red circles. In the upper triangular panels, pairwise correlations are represented by circles whose size and colour intensity reflect their magnitude (see colorbar), computed for the full sample and for the subsets of regular units and outliers. For example, for ROBBERY and SEIZED, the correlation is stable for the full sample and regular units, but weak and slightly negative among outliers.

5. Detailed description of a representative aggregation step

As illustrated in the previous section, the proposed hierarchical framework is articulated through a sequence of intermediate aggregation steps, each aimed at synthesizing homogeneous groups of variables into sectoral indicators. In this section, we provide a detailed illustration of a single aggregation step, with the purpose of clarifying both the robust statistical methodology adopted and the role of interactive graphical tools in supporting interpretation and decision-making.

5.1. Reference set of indicators

As an illustrative example, we consider an aggregation step based on a subset of indicators describing different dimensions of criminal activity and economic illegality. Specifically, the following variables are included:

- Crimes against the person (`PERSON`);
- Thefts (`THEFTS`);
- Robberies (`ROBBERY`);
- Organized crime and financial/economic crimes (`ORGFNC`);
- Crimes against public safety (`PUBSAF`);
- Other crimes (`OTHERS`);
- Assets seized from organized crime and anti-mafia measures (`SEIZED`);
- Undeclared (shadow) economy (`SHADOW`).

The starting data matrix has dimension 106 × 8. All variables are preliminarily standardized in order to remove scale effects and to ensure comparability.

5.2. Detection of multivariate outliers

The first step of the analysis consists in the identification of multivariate outlying units. Either the MCD [27] or the forward search [28] can be used to detect the atypical observations. Given that the procedure is used multiple times it is important to use a Bonferronized confidence level with a very high confidence level in order to identify and exclude only clearly atypical units [29]. It is important to remark that the atypical units are just removed from the calculation of PCA but their score is computed. Fig. 2 reports a scatterplot matrix of the selected variables, where observations identified as outliers are highlighted. The variables are shown in jittered and standardized form to improve readability and keep confidentiality.

The univariate boxplots displayed on the main diagonal indicate that, for most variables, the distribution of the non-outlying units is markedly right-skewed. In the upper triangular panels, circles proportional in size and colour intensity to the magnitude of the pairwise correlation coefficients are shown. Distinct graphical representations are used for correlations computed on the full sample and on the subsets of regular and outlying units. This visualization highlights the impact of atypical observations on the dependence structure among variables and supports the use of a robust approach. In this example, excluding outliers leads to a noticeable increase in the estimated correlations.

5.3. Robust PCA and interactive biplot

Fig. 3 displays the interactive bivariate biplot obtained from robust PCA. The left panel refers to the analysis conducted on the full dataset, while the right panel reports the results after removing the detected outliers. In both cases, a 95% confidence ellipse is superimposed.

The interactive biplot allows the user to:

- show or hide the labels of the territorial units;
- colour observations according to their orthogonal distance from the subspace spanned by the retained components;
- control the proportion of units displayed;
- change the sign of the components to enhance interpretability.

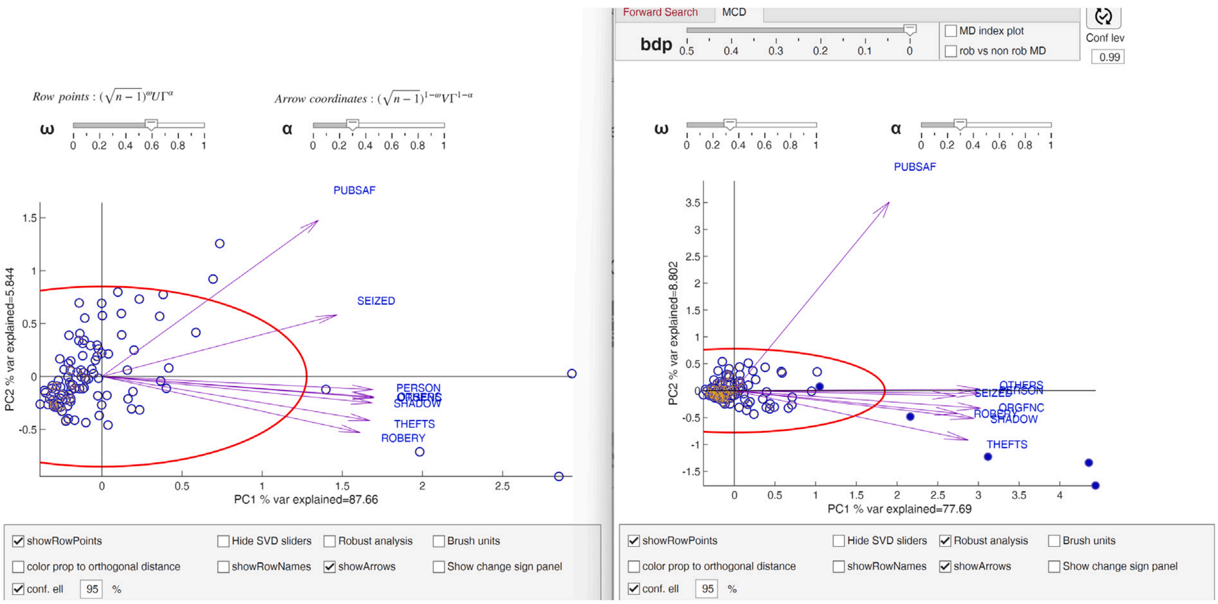

Fig. 3. Interactive bivariate biplot based on all units (left panel) and after removal of multivariate outliers (right panel).

The proportion of explained variance indicates that the first principal component captures a substantial share of the overall variability, justifying its use as a synthetic indicator at this aggregation level.

The two sliders displayed in the upper panel of the interactive interface correspond to the parameters α and ω in the generalized singular value decomposition,

$$\mathbf{Z} \approx \left\{(\sqrt{n-1})^{\omega}\mathbf{U}_{(2)}\boldsymbol{\Gamma}^{\alpha}_{(2)}\right\}\left\{\boldsymbol{\Gamma}^{1-\alpha}_{(2)}\mathbf{V}^{\top}_{(2)}(\sqrt{n-1})^{1-\omega}\right\},$$

which governs the relative emphasis placed on scores and loadings in the biplot representation.

The influence of outliers is evident. When all units are included, the loading associated with the variable SEIZED appears intermediate between PUBSAF and the remaining variables. After removing the outliers, SEIZED emerges as strongly correlated with most of the other indicators, while PUBSAF shows a comparatively weaker association. Moreover, two provinces clearly aligned with the direction of PUBSAF in the full-sample analysis substantially change their position once outliers are excluded. The confidence ellipse also becomes more elongated, indicating a clearer underlying structure.

5.4. Interactivity and geographical interpretation

A key feature of the proposed framework is the tight integration between multivariate analysis and interactive visualization tools. Through the interactive biplot, analysts can brush selected regions of the score space and immediately visualize the corresponding territorial units in the scatterplot matrix, and the geographic maps (Fig. 4). This functionality is particularly valuable for operational tasks, as it enables the rapid identification of geographically clustered behaviours and facilitates the interpretation of complex multivariate patterns.

Finally, it is worth noting that, in this specific aggregation step, all elements of the first robust eigenvector are positive. Consequently, there was no need to enforce the non-negativity constraint on the loadings introduced in Section 4.

6. Conclusions

This paper has proposed a coherent, robust, and interpretable framework for the construction of composite indicators aimed at supporting strategic decision-making in Public Administrations, with a specific application to the Italian *Guardia di Finanza*. In a context characterized by increasing institutional complexity and growing constraints on available resources, the need for statistically sound tools capable of synthesizing large volumes of heterogeneous information is particularly pressing.

The proposed methodology is based on a hierarchical aggregation scheme in which constrained robust principal component analysis is systematically applied at each stage. This choice allows dimensionality reduction while limiting the influence of atypical observations, which are common in socio-economic and crime-related data. By avoiding black-box procedures and preserving the linear structure of the model, the framework ensures a high degree of transparency and interpretability, enabling analysts to trace the contribution of each original variable to the final composite indicator.

A key feature of the approach is the integration of heterogeneous information sources. External indicators describing institutional, social, and economic conditions are combined with internal data collected directly by the *Guardia di Finanza*, which provide a unique observational perspective on economic-financial and organized crime phenomena. The availability of three parallel indicator systems, based on external data only, internal data only, and the integrated dataset, makes it possible to assess the stability of territorial rankings and to identify provinces whose complexity is not fully captured by traditional statistical sources.

The flexibility of the aggregation process further represents a major strength of the proposed framework. Alternative weighting schemes can be adopted to account for the proportion of variance explained by sectoral indicators and for the different number of variables contributing to each domain. This flexibility supports extensive sensitivity analyses and allows the institution to tailor the construction of composite indicators to specific analytical or operational objectives.

From an operational standpoint, the linear representation of the final index facilitates both the interpretation of results and the evaluation of territorial contributions. Each province can be positioned within a multidimensional analytical space whose structure is directly linked to observable variables, and whose geometry can be explored through interactive graphical tools. This enhances the usability of the indicators by non-technical users and supports informed decision-making in resource allocation and operational planning.

Overall, the proposed framework provides a statistically rigorous yet practically oriented solution for measuring socio-economic and

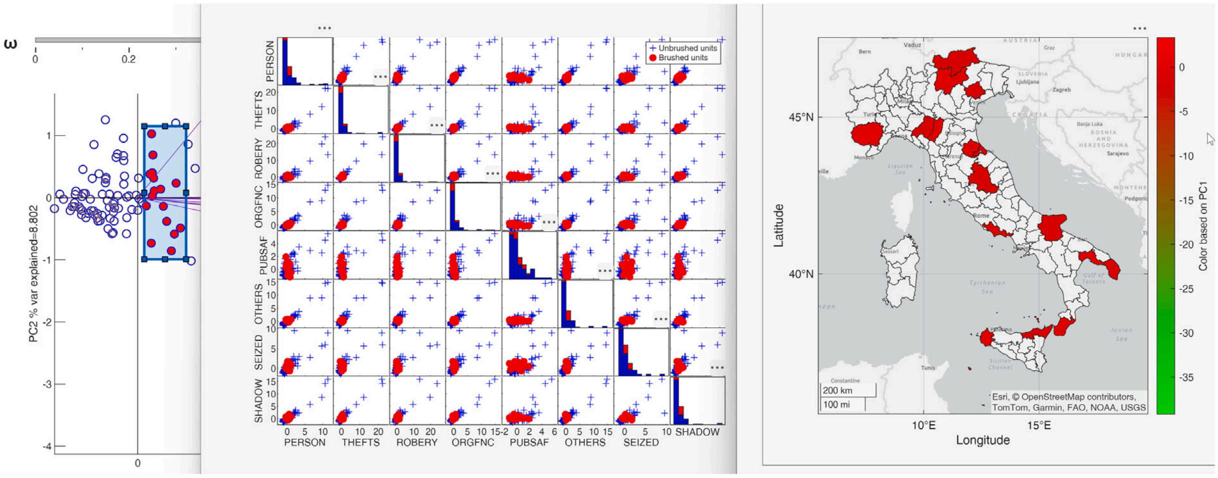

Fig. 4. Interactive dynamic biplot with linked views. Brushing a region of the score space automatically highlights the corresponding units in the scatterplot matrix and the geoplot. This figure is provided for illustrative purposes only, showing how the GDF can at any moment, through a simple brushing action, retrieve additional information about the units occupying a selected region; in this example, the highlighting is completely unrelated to any ranking of the units, and the red colour simply indicates the brushed units.

criminal complexity at the territorial level. While the application presented here focuses on the Italian context, the methodology is general and can be readily adapted to other institutional settings where heterogeneous data sources and robustness requirements play a central role. Future developments may include the extension of the framework to dynamic analyses over time and the integration of additional interactive features to further enhance its operational value.

CRediT authorship contribution statement

Marco Tolla: Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Investigation, Data curation. **Marco Riani:** Writing – original draft, Software, Methodology, Conceptualization. **Paolo Mariani:** Writing – original draft, Validation, Supervision, Methodology, Investigation, Conceptualization. **Corrado Crocetta:** Writing – original draft, Validation, Supervision, Methodology, Investigation, Conceptualization.

Acknowledgements

The present work is part of the cooperation between Guardia di Finanza (GdF, Italian Economic and Financial police), Italian Statistical Society (SIS), University of Milan Bicocca and University of Parma. We wish to thank the Guardia di Finanza General Headquarters; the Research Centres, Ro.S.A. of the University of Parma and the B-ASC of the University of Milan Bicocca.

Appendix

The implementation of the algorithm used to compute the synthetic indicators relies on the FSDA (Flexible Statistics and Data Analysis) toolbox for MATLAB. This toolbox, developed by the University of Parma in collaboration with the Joint Research Centre of the European Commission, provides a comprehensive set of advanced statistical methods with a particular focus on robust analysis. FSDA is publicly available and can be freely downloaded from the MathWorks FileExchange. The MATLAB code required to replicate the results in Figs. 2 and 3 is available in a dedicated GitHub repository at the following address: https://github.com/MarcoRianiUNIPR/TRMC

Data availability

The data that has been used is confidential.

References

[1] Sellin T. The basis of a crime index. J Crim Law Criminol 1931;22(3):335–56.
[2] Blumstein A. Seriousness weights in an index of crime. Am Sociol Rev 1974;39(6):854–64.
[3] Parker RN, McDowall D. Constructing an index of officially recorded crime: The use of confirmatory factor analysis. J Quant Criminol 1986;2(3):237–50.
[4] Parton DA, Hansel M, Stratton MJR. Measuring crime seriousness: Lessons from the national survey of crime severity. Br J Criminol 1991;31(1):72–85.
[5] MacDonald Z. Official crime statistics: Their use and interpretation. Econ J 2002;112(477):F85–106.
[6] Cesario E, Lindia P, Vinci A. Multi-density crime predictor: an approach to forecast criminal activities in multi-density crime hotspots. J Big Data 2024;11(1):75.
[7] Maggino F. La misurazione dei fenomeni sociali attraverso indicatori statistici. Aspetti metodologici. Firenze: Firenze University Press; 2009.
[8] Mazziotta M, Pareto A. Gli indici sintetici. Torino: Giappichelli; 2020.
[9] Gismondi R, Russo MA. Synthesis of statistical indicators to evaluate quality of life in the Italian provinces. Stat Appl 2008;6(1):33–56.
[10] Jolliffe IT. Principal component analysis. 2nd ed.. New York: Springer; 2002.
[11] Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev: Comput Stat 2010;2(4):433–59.
[12] Abdi H, Williams LJ, Valentin D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. WIREs Comput Stat 2013;5(2):149–79.
[13] OECD. Handbook on constructing composite indicators: Methodology and user guide. Paris: OECD Publishing; 2008.
[14] Parisi G. La chiave, la luce e l’ubriaco. come si muove la ricerca scientifica. Rome: Di Renzo Editore; 2021.
[15] Parisi G. In a flight of starlings: The wonders of complex systems. London: Allen Lane; 2023.
[16] Hubert M, Rousseeuw PJ, Vanden Branden K. Robust principal component analysis. Technometrics 2005;47(1):64–79.
[17] Schuberth F, Henseler J, Dijkstra TK. Partial least squares path modeling using ordinal categorical indicators. Qual Quant 2020;54:1345–63.
[18] Henseler J, Dijkstra TK. Consistent partial least squares path modeling. MIS Q 2015;39(2):297–316.
[19] Crocetta C, Ciccarelli AFG, D’Esposito MR. Higher-order composite indicators for university performance evaluation: A PLS-PM approach. Soc Indic Res 2021;156:1021–45.
[20] Mazziotta M, Pareto A. The Mazziotta–Pareto index: Advances and applications in composite indicator construction. Soc Indic Res 2026. Forthcoming.
[21] Brüggemann R, Patil GP. Ranking and prioritization for multi-indicator systems: Introduction to partial order applications. New York: Springer; 2011.
[22] Eurostat. Commission delegated regulation (EU) 2019/1755. 2019, In force from 1 January 2021.
[23] ISTAT. Codici delle unità amministrative. 2026, https://www.istat.it/classificazione/codici-dei-comuni-delle-province-e-delle-regioni/. Last accessed: May 5th 2026.
[24] Escofier B, Pagès J. Analyses factorielles simples et multiples: Objectifs, méthodes, interprétation. Paris: Dunod; 2020.

[25] ISTAT. Sistema informativo sulla qualità. 2026, https://siqual.istat.it/SIQual/welcome.do. Last accessed: May 5th 2026.
[26] Crosato L, Mussini M, Mariani P, Zavanella B. Linking administrative tax records and survey expenditure data at the local level. In: Crescenzi F, Mignani S, editors. Statistical methods and applications from a historical perspective. Studies in theoretical and applied statistics, Springer; 2014.
[27] Rousseeuw PJ, Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. Technometrics 1999;41:212–23.
[28] Riani M, Atkinson AC, Cerioli A. Finding an unknown number of multivariate outliers. J R Stat Soc Ser B Stat Methodol 2009;71(2):447–66.
[29] Riani M, Atkinson AC, Perrotta D. A parametric framework for the comparison of methods of very robust regression. Statist Sci 2014;29:128–43.

Marco Tolla is a General in the Guardia di Finanza (Italian Economic and Financial Police), where he has held several senior operational and strategic roles. His work focuses on economic and financial crime, risk analysis, and the use of statistical and analytical tools to support decision-making and enforcement activities. He has contributed to the development and application of data-driven methodologies within public administration and law enforcement contexts.

Marco Riani is Full Professor of Statistics at the University of Parma, Italy. He holds a Ph.D. in Applied Statistics from the University of Florence. He won the prize for the best Ph.D. thesis in statistics. His research interests include robust statistics, regression, multivariate analysis, time series, and statistical methods for complex data, with applications to economic analysis, fraud detection, and decision support. He is coauthor of the MATLAB toolbox FSDA and has published extensively in international statistical journals. He has led and contributed to several national and European research projects and currently serves as Associate Editor for Metron, Statistical Methods and Applications, and STATS.

Paolo Mariani is Full Professor of Statistics at the University of Milano-Bicocca, Italy. His research interests include statistical methods for official statistics, composite indicators, data quality, and multivariate analysis, with applications in economic and social measurement. He has published in international journals and has been involved in several national and international research projects, collaborating with public institutions and statistical agencies.

Corrado Crocetta is Full Professor of Statistics at the University of Bari Aldo Moro, Italy. His research focuses on applied statistics, data analysis, and statistical methods for economic and social sciences, with particular attention to methodological developments and real-world applications. He has authored numerous scientific publications and has participated in national and international research initiatives.