MULTIVARIATE ANALYSIS FOR EXPLORING SUB-MUNICIPAL DEPRIVATION OF ITALIAN MUNICIPALITIES[1]

Samanta Pietropaoli

Abstract. This study presents a multivariate analysis of nine individual indicators representing socio-economic deprivation at the sub-municipal level, based on data from 25 Italian municipalities. The aim is to investigate the internal heterogeneity of urban areas by examining the statistical properties and interrelationships of indicators reflecting different dimensions of local deprivation.

The first phase explores the distributional characteristics of the individual indicators, highlighting asymmetries, heavy tails, and outliers. Particular attention is given to extreme values, which may reveal critical conditions in specific urban zones and influence the robustness of conventional statistical summaries.

Principal Component Analysis is then applied to synthesise the complexity of the information, allowing dimensionality reduction and the identification of latent structures underlying the data. Findings highlight recurring patterns of socio-economic disadvantage common to several urban contexts, as well as context-specific vulnerabilities pointing to shared structural issues.

The ultimate goal is to provide a statistical rationale for selecting the individual indicators used to compute the composite sub-municipal deprivation index. At the same time, the study offers a concise and interpretable representation of spatial disparities, with potential to inform local policy and planning interventions.

1. Introduction

Urban areas are increasingly characterised by pronounced internal disparities in socio-economic conditions, often masked by aggregate municipal-level indicators. Understanding sub-municipal variations is essential for developing targeted policies to address local vulnerabilities and promote inclusive development. Statistical methods capable of capturing and synthesising this complexity are therefore crucial.

This paper investigates the internal heterogeneity of socio-economic deprivation[2] across 25 Italian municipalities by applying Principal Component Analysis (PCA)

---

[1] This article exclusively reflects the author's opinions.

[2] Socio- economic deprivation is defined as a condition in which families and individuals experience difficulties in adequately meeting their basic needs due to a lack of or insufficient economic, employment, educational, and social resources and opportunities. (Carbonetti *et al*., 2025)

to nine individual indicators. These indicators form the empirical basis for the construction of ISTAT's sub-municipal composite deprivation index, enabling identification of recurrent disadvantage patterns and context-specific phenomena (Carbonetti *et al.*, 2025). The sample includes cities from northern, central, southern, and insular Italy, representing diverse urban and socio-economic contexts. Building on previous PCA applications in deprivation measurement (e.g., Messer *et al.*, 2006; Basu and Das, 2021), this study explores the latent multidimensional structure of deprivation at the sub-municipal level, addressing a methodological gap in the existing literature, which has predominantly focused on broader spatial scales. The contribution of this paper is twofold: first, it critically examines the statistical properties of the indicators—focusing on asymmetry, outliers and heavy-tailed distributions—and second, it employs PCA to explore latent deprivation dimensions, assessing whether one or more underlying dimensions best capture the phenomenon.

This work is part of a broader institutional effort to improve local socio-economic vulnerability measurement through integration of census and administrative data. It aims to support spatially targeted interventions and foster more granular territorial diagnostics. The remainder of the paper is structured as follows: Section 2 reviews the literature on PCA in similar contexts. Section 3 presents the indicators. Section 4 examines their distributional patterns. Section 5 reports and interprets PCA results. Finally, Section 6 concludes.

2. Review of the Literature

PCA (Pearson, 1901) is a multivariate statistical technique widely used in constructing composite indicators, reducing data dimensionality by summarising the variance of correlated variables into a smaller set of uncorrelated components (Ram, 1982). The first component captures the maximum variance, followed by subsequent components explaining decreasing residual variance. PCA has been applied in well-being and deprivation indices, particularly to derive data-driven weighting schemes for the components of a composite indicator. Notable applications include Ram's Physical Quality-of-Life Index, Noorbakhsh's adaptation for the Human Development Index (1998), and contributions by Klasen (2000), McGillivray (2005), as well as OECD and European Commission initiatives (Saisana and Tarantola, 2002; Bandura, 2008). Standard practice often involves using the first principal component's factor loadings as weights for a composite indicator (Greyling and Tregenna, 2016).

However, PCA has well-recognised limitations, it may produce negative weights or yield components that are difficult to interpret when indicator correlations are weak (Saisana and Tarantola, 2002). PCA is primarily an exploratory tool, useful for identifying patterns, detecting redundancy or understanding underlying structures

(Ogwang and Abdou, 2003). Recent studies emphasise its value in revealing multi-dimensional constructs, where multiple uncorrelated dimensions represent distinct latent phenomena (Terzi *et al.*, 2021). In this sense, PCA is particularly suited to identifying redundant variables and uncovering whether a multidimensional concept—such as a deprivation index—consists of several uncorrelated dimensions, each potentially representing a distinct latent factor. When such distinct dimensions emerge, PCA proves especially valuable in revealing the internal structure of the data. In these cases, multidimensional approaches may be preferable to traditional aggregation methods based on predefined weights. However, as Terzi *et al.* (2021) also warn, PCA may not be ideal for final construction of composite indices, since the orthogonality it imposes among components does not necessarily reflect real-world interdependencies across dimensions.

3. Indicators

The empirical analysis draws on data from 25 Italian municipalities selected to capture a broad spectrum of urban contexts and socio-economic configurations across the country. These municipalities—Turin, Genoa, Milan, Verona, Venice, Padua, Gorizia, Trieste, Parma, Modena, Bologna, Florence, Perugia, Rome, Prato, Carpi, Naples, Bari, Taranto, Reggio Calabria, Palermo, Messina, Catania, Olbia, and Cagliari—are geographically distributed throughout Italy, covering northern, central, southern, and insular regions, and include both large metropolitan areas and medium-sized urban centres. The analysis focuses on enumeration areas[3] (EAs), the smallest statistical units, providing high spatial resolution for intra-urban disparities. The analytical framework enhances robustness by filtering only EAs with a stable residential presence and clear urban characteristics—such as buildings, infrastructure, and street networks—while excluding zones[4] with negligible population or atypical land use (e.g., parks, cemeteries, hospitals).

The number of EAs per municipality ranges from 272 in Gorizia to 11,688 in Rome, with an average of 2,180 and a median of 1,476[5]. All EAs were treated as equally weighted observational units—a one-to-one strategy—to capture the relative

[3] According to ISTAT (2015), an enumeration area is the smallest territorial unit used for data collection within a municipality, primarily designed to organise census operations. It generally corresponds to a city block or part of it. These fine-grained spatial units are essential for capturing local socio-economic information and are also used for geocoding processes and delineating functional areas.
[4] This exclusion was limited to specific zones within the cities of Florence, Bologna, Milan, Rome, and Turin.
[5] The number of Enumeration Areas (EAs) varies across cities due to differences in city size and population, as well as the selection of EAs meeting completeness criteria. The proportion of EAs included relative to the total number per city varies slightly, but this does not introduce bias in the subsequent analysis.

spatial structure of deprivation rather than to produce population-weighted estimates, which could have overemphasised larger or more densely populated units. Since the indicators were expressed in relative or per-capita terms, differences in population size were already accounted for, ensuring meaningful comparability across units. Indicators were derived from administrative sources and sample surveys—primarily from ISTAT, but also from the National Social Security Institute (INPS), the Ministry of Education (MIM), the Ministry of University and Research (MUR), and the Cadastre Register—using the most recent sub-municipal data available (2020–2021, depending on the source).

These are grouped into three thematic domains: economic hardship, labour market exclusion, and educational disadvantage. Priority was given to indicators capturing observable and measurable conditions rather than potential risk factors, ensuring a tangible and measurable representation of local deprivation. The nine indicators were selected through a collaborative expert-based process, which identified the most appropriate measures based on data availability and conceptual relevance. This study represents an experimental application developed within the methodological framework proposed by Carbonetti et al. (2025), with some adjustments to the indicator set reflecting data constraints and contextual refinements. The formative measurement model (Mazziotta and Pareto, 2019) supports this expert-driven and context-sensitive approach to indicator selection. Particular attention was paid to statistical quality, including relevance, accuracy, and comparability across municipalities. This structured selection process ensured theoretical coherence and empirical robustness, providing a solid foundation for the construction of the composite deprivation index (Carbonetti et al., 2025). Altogether, the selected indicators offer a nuanced depiction of territorial fragility at the sub-municipal level, as summarised in Table 1.

All indicators are oriented so that higher values correspond to greater deprivation, except for DISOCC_1 (employment rate), where higher values indicate better labour integration. Each indicator was computed at the EA level and standardised (z-scores)6 prior to analysis.

[6] Although PCA is sensitive to outliers, z-score standardization was adopted to preserve the relative variability and comparability of indicators across municipalities. Outliers were considered an intrinsic feature of spatial heterogeneity rather than noise. A more robust scaling, such as the Median Absolute Deviation, could have dampened these meaningful territorial contrasts.

Table 1 – *List of individual indicators used for the construction of the sub-municipal deprivation index.*

Indicator	Description	Prevalent Component	Sources
DISECO_1	% of the population aged 70 and over living alone and without owning a home	Economic	PPHC[7], Cadastre Register
DISECO_2	% of individuals in households where no member is employed or receives a work-related pension	Economic	PPHC, Pensioners' records
DISECO_3	Individuals in households at risk of poverty per 100 residents.	Economic	PPHC, Income Register
DISOCC_1	% of employed individuals aged 25–64 over the total population aged 25–64	Employment	PPHC
DISOCC_2	% of individuals aged 0 to 64 living in households with very low work intensity	Employment	PPHC, Labour Register
DISOCC_3	% of employed individuals (aged 25 to 64) with predominantly precarious employment conditions during the year, over the total number of employed individuals (25–64) in the year	Employment	PPHC, Labour Register
DISEDU_1	% of the population aged 25 to 64 without upper secondary school education	Educational	PPHC
DISEDU_2	% of individuals aged 15–29 who are not employed and not enrolled in any regular education program of the MUR or the MIM	Educational	PPHC, MIM, MUR
DISEDU_3	% of students who drop out or repeat the year (Proxy indicator of potential dropout in secondary school)	Educational	PPHC, Education Register

Sources: ISTAT, INPS, Cadastre, MIM, MUR (2020-2021).

4. Distributional Patterns and Variability of Indicators

As a first step, the distributional characteristics of the nine indicators were analysed, with skewness and kurtosis statistics computed to detect heavy tails and outliers that could distort subsequent multivariate analysis. This exploratory phase revealed substantial heterogeneity in shape and dispersion across the 25 municipalities.

Large urban centres such as Rome, Naples, and Palermo display pronounced asymmetries and extreme values across most DISECO and DISOCC indicators. For instance, DISECO_1 (elderly living alone without home ownership) shows a long right tail, while DISEDU_2 (NEET[8] rate) exhibits strong positive skewness, indicating spatially concentrated vulnerability and internal segmentation. Outliers are fre-

[7] *PPHC* stands for *Permanent Population and Housing Census.*

[8] NEET (Not in Education, Employment, or Training) refers to young people, typically between the ages of 15 and 29, who are not in education, employment, or training. This means they are neither enrolled in formal education or training programs nor engaged in paid or unpaid work.

quent in these metropolitan areas, with some EAs diverging sharply from the surrounding urban fabric, indicating neighbourhoods where poverty, low employment, or dropout rates reach extreme levels not observed elsewhere in the same city.

In contrast, cities like Parma, Padua, and Modena show more symmetric and compact distributions, with lower dispersion and fewer anomalies. Indicators such as DISOCC_1 (employment rate) and DISEDU_1 (low education) suggest relatively uniform conditions and less spatial concentration of deprivation, supporting the hypothesis of a more homogeneous socio-territorial structure. Some municipalities exhibit irregular or multimodal distributions. For example, DISEDU_3 (students at risk of dropout or repetition) presents a bimodal pattern in Cagliari and Olbia, possibly reflecting contrasts between central and peripheral areas or between resident and seasonal populations. In northern cities such as Trieste and Gorizia, age-related vulnerability (DISECO_1) is particularly evident, with right-skewed distributions highlighting the challenges of ageing populations in housing-insecure settings. At the same time, smaller cities such as Reggio Calabria and Taranto, also show fat-tailed distributions in DISOCC_2 and DISEDU_3, indicating that acute localised disadvantage can emerge even in medium-sized urban contexts. Outliers in these areas often signalling deeply marginalised neighbourhoods. A notable pattern is observed for DISOCC_3 (precarious employment), which exhibits high variability in both central and peripheral zones. In cities like Naples and Rome, this reflects dual labour markets with the coexistence of widespread precariousness and pockets of stable employment. Educational deprivation, particularly DISEDU_2 and DISEDU_3, is unevenly distributed across the country: right-skewed distributions reveal extreme vulnerability in Palermo, Taranto, and Reggio Calabria, whereas northern cities such as Bologna and Milan show tighter and more balanced profiles.

Overall, the presence of skewed and heavy-tailed distributions, combined with significant outliers, confirms the complexity of socio-spatial vulnerability in Italian cities. These findings underscore the importance of developing a composite index at the sub-municipal level, designed to capture localised disparities and distributional patterns that would otherwise be obscured in aggregated analyses.

5. Principal Component Analysis: selected results

Once the dataset was standardised, preliminary checks were conducted to assess the suitability of the indicators for PCA. Two commonly used diagnostic tests were applied: the Kaiser-Meyer-Olkin (KMO)[9] measure of sampling adequacy and Bartlett's Test of Sphericity (BTS)[10]. All municipalities passed these preliminary tests, with KMO values exceeding the 0.70 threshold, indicating moderate to excellent sampling adequacy. The highest KMO value was observed in Naples (0.88), reflecting excellent adequacy, followed by Palermo (0.85), Cagliari (0.82), and both Rome and Trieste (0.78). Regarding the BTS test, the results confirmed that the data were suitable for the analysis.

The PCA was conducted using the *R* statistical software, specifically the `princomp()` and `PCA()` functions from the stats and FactoMineR packages. The analysis was based on the correlation matrix of standardised indicators. No rotation was applied, as the aim was to explore the overall structure of the data and identify the principal latent dimensions explaining the maximum variance, rather than to obtain a simpler or more interpretable factor pattern. Negative loadings were interpreted as variables inversely associated with the corresponding component, meaning that higher values of these variables reflect lower scores on that latent dimension. Although PCA is sensitive to sample size, the use of relative and per-capita indicators ensures comparability across cities, maintaining the robustness of the results even in smaller cities with fewer EAs.

The PCA results reveal two dominant components explaining a substantial share of total variance in most municipalities, although in some cases a third component is required to capture socio-economic and educational complexity. This is further illustrated by correlation circles, which visually represent the relationships between indicators and the first two principal components. The direction and length of the arrows indicate the contribution and strength of each indicator in the component space, supporting the interpretation of latent structures and enabling cross-city comparison of dominant dimensions.

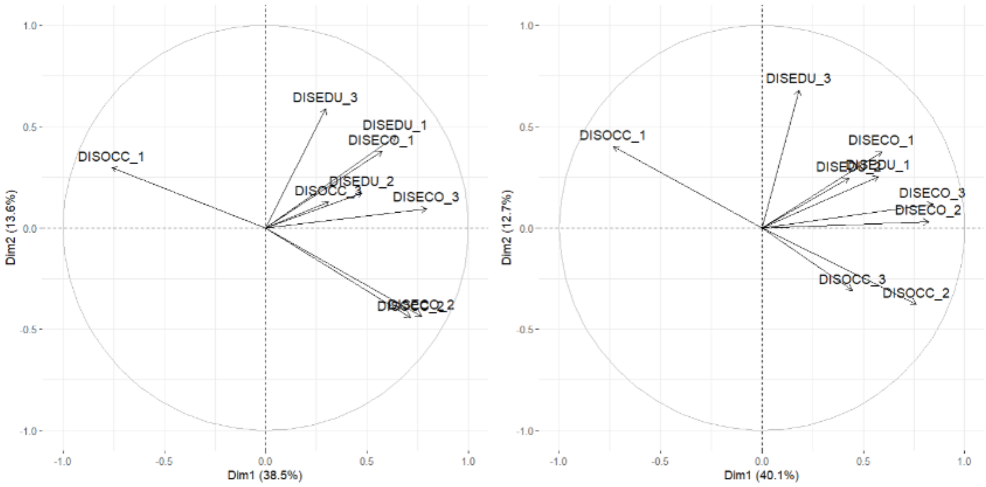
To explore territorial differences more concretely, the results for selected pairs of cities—representative of the North, Centre, South, and Islands—are presented below. These examples highlight how the relative contribution of each indicator to the principal components varies across contexts, revealing both recurring structures and local specificities in patterns of deprivation. Figure 1 shows the PCA correlation circles for Milan and Trieste, where the first two components explain around 52%

[9] The KMO statistic evaluates the proportion of variance among variables that might be common variance. Its value ranges between 0 and 1, with higher values indicating a greater degree of shared variance and therefore better suitability for factor-based techniques. While a KMO value above 0.6 is generally considered acceptable, values above 0.7 or 0.8 are preferable and indicate good to excellent adequacy (Kaiser, 1974).

[10] The BTS assesses whether the correlation matrix significantly differs from an identity matrix—that is, a matrix in which all diagonal elements are equal to 1 and all off-diagonal elements are 0. If the test is statistically significant ($p < 0.05$), the null hypothesis that the variables are uncorrelated can be rejected, indicating that the data are appropriate for structure detection through PCA (Bartlett, 1951).

and 53% of the variance, respectively. In both cities, PC1 captures structural socio-economic vulnerability, with high positive loadings on DISECO_2 and DISECO_3 (economic hardship) and strong negative loadings on DISOCC_1 (employment rate). Educational exclusion also contributes to PC1, more markedly in Milan through DISEDU_1 and DISEDU_2. PC2 differentiates educational dynamics with DIS-EDU_3 (youth dropout risk) emerging as a major contributor, especially in Milan, where educational deprivation aligns with both PC1 and PC2. Trieste shows a similar, though less pronounced, pattern.

Figure 1 – *Correlation Circles from PCA: Milan (left) vs. Trieste (right): PC1 (Economic and Labour Vulnerability) vs. PC2 (Educational Deprivation)***.**



In Rome and Florence (Figure 2), PC1 explains 38–40% of the variance and reflects economic hardship and labour exclusion, with strong contributions from DIS-ECO_2, DISECO_3, and DISOCC_2, and a negative loading on DISOCC_1. Educational indicators contribute moderately to PC1 in both cities. PC2, explaining around 12–13% of the variance, captures educational exclusion more distinctly in Florence, especially through DISEDU_1 and DISEDU_3. In Rome, PC2 shows weaker or negative correlations with educational indicators, suggesting a less pronounced role. Florence also displays a moderate positive loading of DISECO_1 on PC2, pointing to localised socio-demographic variation.

In Naples and Bari (Figure 3), PC1 dominates the structure, explaining 44–46% of the variance, and reflects structural socio-economic hardship driven by economic dependency (DISECO_2, DISECO_3), low work intensity, and precarious employment (DISOCC_3). Educational indicators contribute more marginally, with variable patterns across components. These results confirm that deprivation in both cities is primarily shaped by entrenched economic and labour vulnerabilities, with local nuances in the interaction between dimensions.

Figure 2 – *Correlation Circles from PCA: Rome (left) vs. Florence (right): PC1 (Economic and Labour Exclusion) vs. PC2 (Educational Exclusion).*

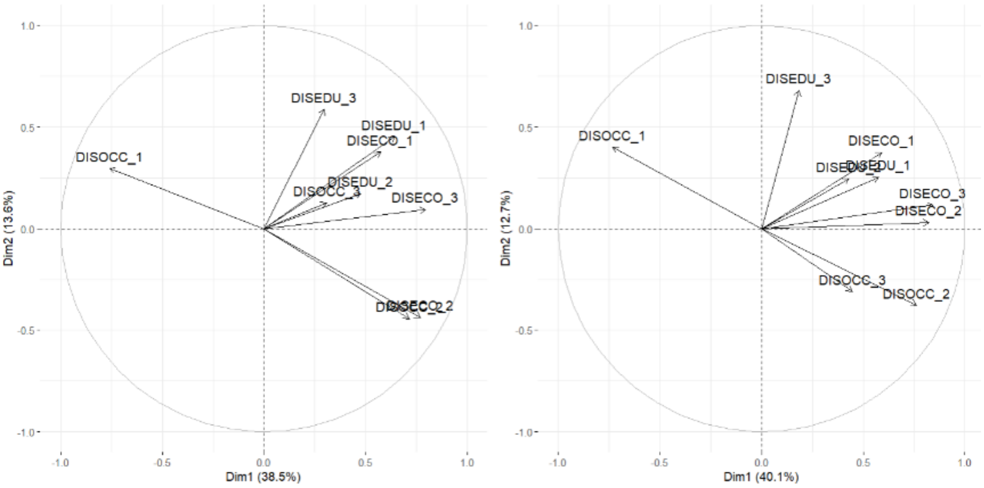


Figure 3 – *Correlation Circles from PCA: Naples (left) vs. Bari (right): PC1 (Economic and Labour Vulnerability) vs. PC2 (Marginal Educational Variation).*

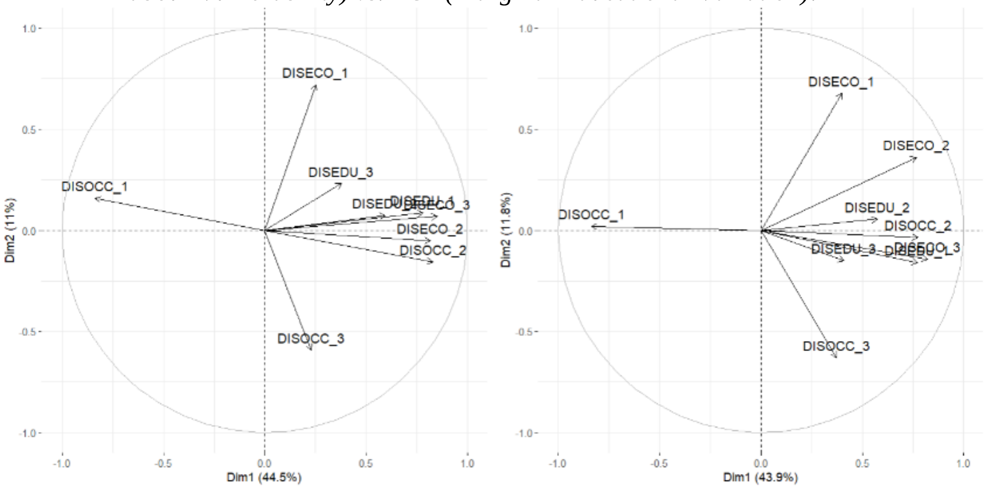
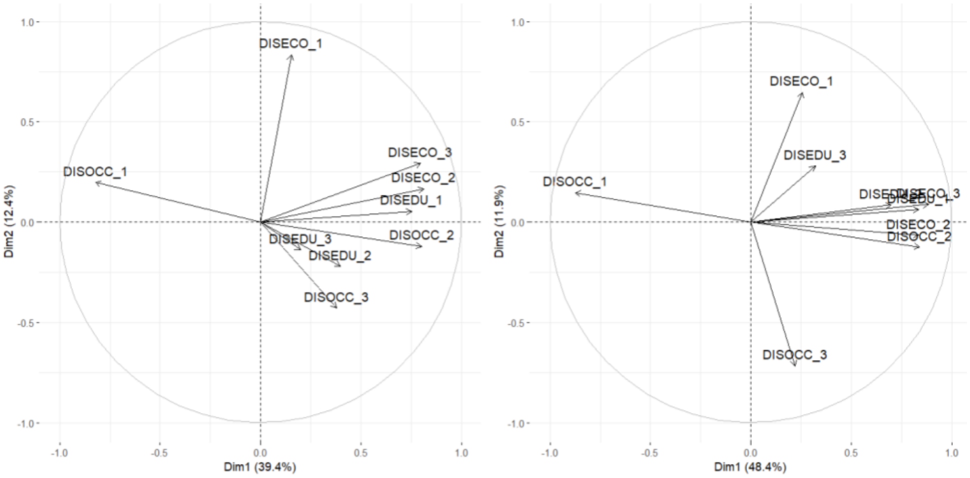


Figure 4 – *Correlation Circles from PCA: Palermo (left) vs. Cagliari (right): PC1 (Economic and Educational Deprivation) vs. PC2 (Labour and Context-Specific Dimensions).*



6. Conclusions

This study has highlighted the importance of sub-municipal analysis for capturing the internal complexity of socio-economic deprivation in Italian cities. By applying PCA to a set of nine indicators across 25 municipalities, it was possible to identify both recurring patterns of disadvantage and spatially concentrated vulnerabilities that would remain obscured by aggregate measures. The distributional analysis revealed skewed indicators, heavy tails, and outliers—underscoring the limits of aggregate statistics in representing intra-urban disparities and reinforcing the need for disaggregated, multidimensional tools. Rather than using PCA as a weighting mechanism, the study employed it as an exploratory technique to assess the internal coherence of the indicators and to detect latent dimensions of deprivation. This approach highlighted the multidimensionality of the phenomenon—spanning structural socio-economic hardship, labour market exclusion, and educational disengagement—which varies in expression and intensity across different urban contexts. These results underscore that indices constructed at the city level risk overlooking significant intra-urban disparities and therefore call for more flexible frameworks—such as sub-municipal approaches—capable of capturing territorial heterogeneity within municipalities. While certain structural components, particularly economic and labour-related vulnerabilities, appear widely shared, other dimensions like educational deprivation and employment precarity exhibit greater territorial variability, reflecting diverse socio-demographic and economic dynamics. Crucially, PCA also provided statistical evidence supporting the adequacy of the selected individual indicators used for the construction of the composite deprivation index. While expert knowledge is essential in guiding indicator selection, it is not sufficient on its own; empirical validation through statistical analysis is necessary to ensure the robustness and relevance of the indicators chosen. This reinforces the value of combining theoretical expertise with data-driven evidence in the design of multidimensional measurement tools. Nevertheless, some limitations should be acknowledged. The analysis is cross-sectional and may therefore not capture temporal dynamics or evolving spatial patterns of deprivation. Moreover, PCA assumes linear relationships among variables and may not fully represent complex or non-linear interactions underlying socio-economic phenomena. Finally, although the selected indicators cover key dimensions of deprivation, the omission of other potentially relevant variables could limit the comprehensiveness of the analysis. Overall, the findings contribute to the refinement of the composite deprivation index at the sub-municipal level, enabling more accurate and interpretable representations of territorial vulnerability. Such tools are essential for supporting evidence-based, spatially targeted policies aimed at reducing intra-urban inequalities and promoting inclusive development. Future work could extend this framework to additional cities and incorporate longitudinal data to enhance its utility for territorial diagnostics and policy planning.

References

BANDURA R. 2008. A survey of composite indices measuring country performance: 2008 update. *Working Paper, No. 96*. United Nations Development Programme, Office of Development Studies, New York.

BARTLETT M. S. 1951. The effect of standardization on a Chi-square approximation in factor analysis. *Biometrika*, Vol. 38, No. 3, pp. 337-344.

BASU T., DAS A. 2021. Formulation of deprivation index for identification of regional pattern of deprivation in rural India. *Socio-Economic Planning Sciences*, Vol.74, issue C, Article 100924.

CARBONETTI G., BIASCIUCCI F., CUTILLO A., MAZZIOTTA M., QUONDAMSTEFANO V., TAMBURRANO M. T., TRONU D. 2025. Measuring socio-economic deprivation at sub-municipal level through the integration of census and administrative data. *Rivista Italiana di Economia Demografia e Statistica*, Vol. 79, No. 1, pp. 7-18.

GREYLING T., TREGENNA F. 2017. Construction and analysis of a composite quality of life index for a region of South Africa. *Social Indicators Research*, Vol. 131, pp. 887-930.

ISTAT 2015. *5- Le Sezioni di censimento* in Atti del 9° Censimento generale dell'industria e dei servizi e Censimento delle istituzioni non profit. Rome: Istituto Nazionale di Statistica.

KAISER H. F. 1974. An index of factorial simplicity. *Psychometrika,* Vol. 39, No. 1, pp. 31–36.

KLASEN S. 2000. Measuring poverty and deprivation in South Africa. *Review of income and wealth*, Vol. 46, No. 1, pp. 33-58.

MAZZIOTTA M., PARETO A. 2019. Use and Misuse of PCA for measuring well-being. *Social Indicator Research*, No. 142, pp. 451-476.

MCGILLIVRAY M. 2005. Measuring non-economic well-being achievement. *Review of Income and Wealth*, Vol. 51, No. 2, pp. 337-364.

MESSER L. C., LARAIA B. A., KAUFMAN J. S., EYSTER J., HOLZMAN C., CULHANE J., ELO I., BURKE J. G., O'CAM P. 2006.The development of a standardized neighborhood deprivation index. *Journal of urban health,* Vol. 83, No. 6, pp. 1041-1062.

NOORBAKHSH, F. 1998. A modified human development index. *World development*, Vol. 26, No. 3, pp. 517-528.

OGWANG T., ABDOU A. 2003. The choice of principal variables for computing some measures of human well-being. *Social Indicators Research*, No. 64, pp. 139-152.

PEARSON K. 1901. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science,* Vol. 2, No. 11, pp. 559-572.

RAM R. 1982. Composite indices of physical quality of life, basic needs fulfilment, and income: A 'principal component' representation. *Journal of Development Economics,* Vol. 11, No. 2, pp. 227-247.

SAISANA M., TARANTOLA S. 2002. State-of-the-art report on current methodologies and practices for composite indicator development. *Ispra, Italy: European Commission, Joint Research Centre, Technological and Economic Risk Management Unit*. Vol. 214, pp. 4-15.

TERZI S., OTOIU A., GRIMACCIA E., MAZZIOTTA M., PARETO A. 2021. *Open issues in composite indicators: a starting point and a reference on some state-of-the-art issues,* Roma: Roma TrE-Press.

Samanta PIETROPAOLI, Istat, pietropaoli@istat.it