A MACHINE LEARNING APPROACH FOR THE ANALYSIS OF DEMOGRAPHIC FEATURES OF THE SOCIO-ECONOMIC DEPRIVATION AT SUB-MUNICIPAL LEVEL

Matilde Bonelli, Giancarlo Carbonetti, Elena Grimaccia, Debora Tronu

Abstract. This study provides a machine learning analysis of the social and demographic features of territories, associated with different levels of deprivation, measured by the Socio-Economic Deprivation Index (SED-I). The SED-I has been developed by the Italian National Institute of Statistics (Istat) and is aimed at measuring socio-economic and educational deprivation at sub-municipal level. In this paper, we exploit the availability of socio-economic indicators and demographic context variables at Sub-Municipal Areas (SMAs) level to provide useful information for policies aimed at contrasting the economic, social, and educational deprivation as measured by the SED-I. In the context of socio-economic deprivation, territories may be characterised by a variety of interrelated variables: indicators related to the level of education in the area, employment, household structure, demographic composition referring to population density, foreign subpopulations, and age, and finally, social conditions, are employed in the analysis. Principal Component Analysis (PCA) is employed to summarise the information contained in this large number of correlated indicators. This unsupervised learning method allows us to analyse the relationships among these variables and the SED-I, pointing out the differences between two different socio-economic environments such as Trieste and Cagliari. Furthermore, a K-means cluster analysis is employed to identify population groups with similar composition in relation to deprivation and demographic features. In the K-means cluster analysis, Elbow and Silhouette methods are used to choose the optimal number of clusters. The analysis is carried out focusing on Trieste and Cagliari data, to provide a robust application on different socio-economic contests.

1. Introduction

The measurement of socio-economic deprivation at a detailed spatial level is increasingly recognised in the literature as essential for the design of effective and equitable public policies. Spatial analyses conducted on very fine territorial levels allow the identification of localised areas of deprivation that are often hidden by larger aggregates (Allik *et al.*, 2016). High levels of detail enable targeted interventions that address the specific needs of the most vulnerable populations, improving resource allocation and reducing social inequalities. Moreover, granular data increases the ability to monitor policy impacts over time and adjust strategies dynamically. The approach is in line with the growing emphasis on evidence-based and place-based policy making.

This study analyses Sub-Municipal Areas (SMA): areas defined by municipalities for functional, administrative, or statistical purposes. Principal Component Analysis (PCA) is employed to summarise the information contained in many correlated indicators. In the context of socio-economic deprivation, SMAs may be characterised by a variety of interrelated features such as the level of education in the area, employment, households' composition, population's citizenship and age and access to services. This unsupervised learning method allows us to analyse the relationships among these variables and between territorial features and deprivation, measured by the Socio-Economic Deprivation Index (SED-I). The SED-I has been developed by the Italian National Institute of Statistics (Istat) to provide a measure of socio-economic and educational deprivation at sub-municipal level (Carbonetti *et al.*, 2025a; Carbonetti *et al.*, 2025b). Results obtained applying the methodology to Trieste and Cagliari have been compared.

This paper aims at answering the following research questions: first, we aim at identifying how social deprivation, measured by the SED-I, is associated to the main demographic, social and economic features of SMAs; furthermore, we classify SMAs according to these features; and finally, we can identify the most vulnerable areas, where to target policies.

To identify groups of territorial units that share similar socio-economic profiles, a K-means cluster analysis is applied. This unsupervised learning method provides homogeneous but mutually distinct clusters based on their values on the SED-I and related variables. Identifying sub-municipal areas with similar deprivation's patterns can be grouped for tailored policy design, thus improving the efficiency of resource allocation.

Together, these ML techniques allow for a comprehensive, flexible, and evidence-based analysis of socio-economic features of deprived areas. Their combined use enhances the interpretability of complex data structures and provides a solid foundation for informed decision-making at different territorial scales.

Following this brief introduction, the second chapter presents the data used in the study, and more specifically the indicators used and the correlations among demographic features available at local level. In the third chapter, we will detail the statistical and machine learning methods employed to address our research objectives. Chapter four presents and interprets the results, explaining how these findings can inform the development of effective policies that consider the fragmentation of the Italian territory and its diverse needs. Finally, some conclusions are drawn.

2. Data

This study analyses the demographic features of the territories, associated with different levels of deprivation, with a focus on the most vulnerable areas of two municipalities, Trieste, and Cagliari. As part of the project, launched by the Italian National Institute of Statistics (Istat) and aimed at measuring the socio-economic deprivation of individuals and households at the sub-municipal level, the most critical areas are identified using the Socio-Economic Deprivation Index (SED-I), a composite index which measures household and individual deprivation as a tangible condition of hardship, as distinct from simple exposure to risk[1]. The definition of deprivation used is: "*a condition in which households and individuals experience difficulties in adequately meeting their basic needs due to insufficient economic, employment, educational and social resources and opportunities*".

After a careful experimental phase (Carbonetti *et al.*, 2025a; Carbonetti *et al.*, 2025b), the following nine individual indicators were identified as manifest variables composing the SED-I, using census, administrative sources, and thematic registers, to measure – with accurate, integrated, and geocoded data – the different components of deprivation at the sub-municipal level:

- Economic deprivation: % of individuals aged 70 and over living alone and not owning a home; % of individuals in households in which no member is employed or receiving a pension from work; % of individuals in low equivalised income households.
- Occupational deprivation: Employment rate 25-64 years old (with negative polarity); % of individuals aged 0-64 living in households with very low work intensity; % of employed persons aged 25-64 "not stable" during the year.
- Educational deprivation: % of individuals aged 25-64 without upper secondary school education; % of individuals aged 15-29 who are not employed and are not enrolled in any regular course of study; % of students dropping out or repeating the year.

These indicators are normalised and aggregated with equal weight using the Adjusted Mazziotta-Pareto Index (AMPI) methodology, which produces a non-compensatory composite index based on not fully substitutable of individual indicators and on non-linear function (Mazziotta and Pareto, 2016; Mazziotta and Pareto, 2017). Here are the main conceptual aspects: the normalization method (Constrained Min-Max Method) of individual indicators allows the comparison of values of statistical units, both in space and time, against a common reference that

[1] Istat already produces an indicator which measures social and material vulnerability of the population. The Social and Material Vulnerability Index measures the exposure of some population groups to situations of risk, such as uncertainty of their social and economic condition (Istat, 2020).

does not change over time; the aggregation function is an average penalised by a variance function that avoids the “compensation effect” typical of linear functions.

Data normalisation is based on a transformation of the individual indicators with respect to two values, referred to as goalposts, sets a statistical unit of reference equal to 100 at a base year. The values obtained fall roughly in the range (70:130), where 100 represents the reference figure.

In this study, SED-I is calculated for Sub-Municipal Areas (SMAs). As the baseline used in the calculation is the value of the individual indicators calculated at municipal level (set equal to 100), the values of SED-I are not comparable between different municipalities.

To study the socio-demographic characteristics of the population living in the most deprived areas, a specific set of indicators has been employed, deriving data mainly from census results (Table 1).

Table 1 – *Socio-demographic indicators used in the analysis.*

Individual indicator		Sources
4	Population Density (population/area)	PPHC
5	Size of households (population/households)	PPHC
6	Foreigners (% as a share of total population)	PPHC
7	Employment rate (15+ years old)	PPHC
8	Young people (0-24) (% as a share of the total population)	PPHC
9	Elderly (65+) (% as a share of the total population)	PPHC
10	Students (% as a share of the total population)	PPHC
11	M/F ratio	PPHC
12	Youth (0-24)/Elderly (65+) ratio	PPHC
14	Foreigners 65+ years old (% as a share of the total foreigners)	PPHC
15	One-person households (% out of total households)	PPHC
16	Households with 5 or more members (% out of total households)	PPHC
17	University graduates aged 25-64 (% of the population aged 25-64)	PPHC
18	Replacement rate (pop. 20-24/pop. 60-64 ratio)	PPHC
19	Individuals in households benefiting from social exclusion transfers (% as a share of the total population)	PPHC Income Register

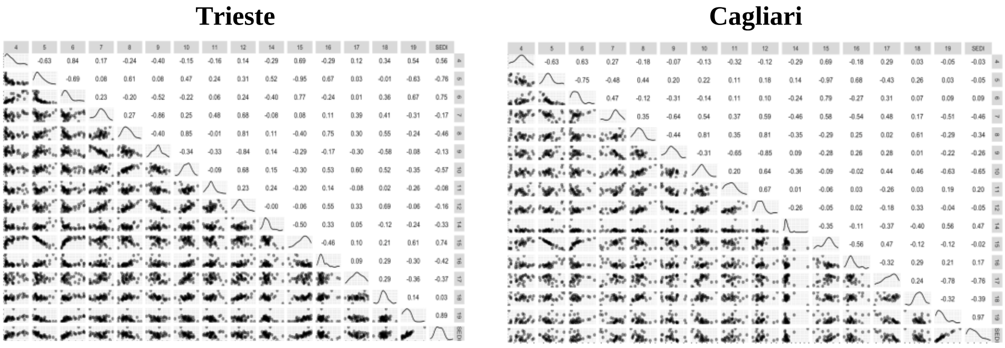
Sources: Istat, Permanent Population and Housing Census (PPHC) and Income Register. Reference year: 2021.

The analysis presented in this paper is focused on Trieste and Cagliari, two municipalities with very different socio-economic conditions and, therefore, needing distinct policies, adapted to their specific context. Both municipalities have a population characterised by a high share of elderly people and one-person households. However, Trieste stands out for its higher number of foreign residents and a more solid employment structure, accompanied by better household economic conditions. In contrast, Cagliari has a lower employment rate and greater economic vulnerability, while its adult population is more educated than that of Trieste.

These differences are confirmed also by the correlation analysis of the indicators used in this study (Figure 1). The correlation matrices of the two municipalities show greater differences in the interrelationships between the employment rate and other socio-demographic variables. In both contexts, employment rates are negatively correlated with the share of elderly population and, to a lesser extent and positive correlated with the proportion of graduates. More marked differences emerge about citizenship and household size. Specifically, in Cagliari, the employment rate shows negative correlations with both the share of families with 5 or more members and the share of foreign residents and positive correlation with a one-person households (as % of total households), whereas such relationships are not observed in Trieste. This suggests that, in Cagliari, lower employment levels are associated with the presence of foreign citizens and larger households, while in Trieste employment levels appear to be mainly influenced by the age structure and the educational attainment of the population. with the proportion of graduates. More marked differences emerge regarding citizenship and household size. Specifically, in Cagliari, the employment rate shows negative correlations with both the share of families with 5 or more members and the share of foreign residents, whereas such relationships are not observed in Trieste. This suggests that, in Cagliari, lower employment levels are associated with the presence of foreign citizens and larger households, while in Trieste employment levels appear to be mainly influenced by the age structure and the educational level of the population.

The analysis of the correlation among SED-I and the other indicators highlights, for both municipalities, a higher positive correlation with individuals in households benefiting from social exclusion transfers. For Trieste, SED-I also shows a strong positive relationship with the share of foreigners and one-person households. For Cagliari, there is a strong negative correlation between SED-I and the share of students (as a percentage of the total population) as well as university graduates aged 25–64 (as a percentage of the population aged 25–64).

Figure 1 – *Correlation of indicators (Istat, 2021).*



Trieste

Cagliari

3. Methods

To achieve the goals of our analysis we have adopted two unsupervised machine learning approaches. The first unsupervised machine learning technique we have employed is the Principal Component Analysis (PCA), which is able to identify a smaller number of representative variables that explain most of the variability of the original dataset, composed of a set of correlated variables (James *et al.*, 2021). We have used this technique to make an unsupervised exploration of the dataset, which includes the SED-I and the set of socio-demographic variables shown in table 1, for the two municipalities under study. The basic idea of performing PCA is that each of our *n* observations lies in a *p*-dimensional space, and we are seeking a smaller number of dimensions to represent them, which are the principal components. We are now interested in knowing the *cumulative proportion of variance explained* (PVE) by the first M principal components (PC), determined by the procedure. For the purpose of PCA we use this quantity to decide how many principal components to retain among all those resulting from the totals, which in general are the min(n-1, p) considering our matrix X with dimension (n x p). A direct tool that we have used is the analysis of the scree plot, where we look for the point at which the curve connecting the PVE across each subsequent PC creates an elbow.

The K-Means cluster method is framed among the unsupervised machine learning techniques and has the aim of finding partitions of the dataset under study with similar values, which are called clusters. Silhouette and Elbow methods have been used to find the optimal number of clusters. The K-means method involves the computation of the within-cluster variation, which summed over all K-clusters must be as small as possible (James *et al.*, 2021). In this paper, we employ the K-Means R-Function, using the Hartigan-Wong algorithm (1979), with squared Euclidean distance to measure similarity between observations. By default, it starts by randomly selecting a set of observations as the initial cluster centres and then iteratively reassigns points and updates centres to minimize the within-cluster sum of squares. This process continues until no further improvement is possible, producing the final cluster assignments and centroids.

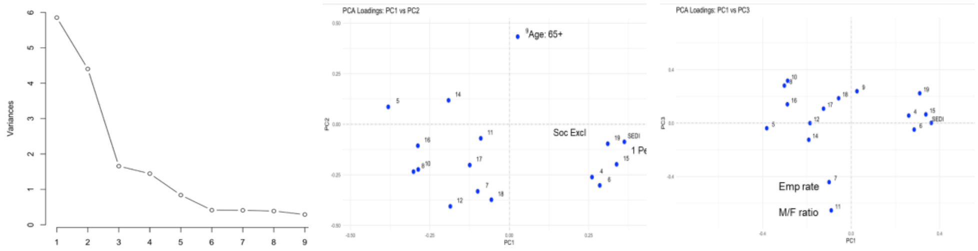
4. Results

To predict the Socio-Economic Deprivation Index (SED-I) in the municipality of Trieste and Cagliari, based on our set of demographic contextual features, we have applied machine learning techniques, as presented in the previous Section. Here, we illustrate which demographic characteristics are most related to the SED-I according to our study. For both municipalities, the SED-I and the set of demographic contextual variables have been included in the computation, after being standardized.

Results of the PCA show that the "Individuals in households benefiting from social exclusion transfers (as a share of the total population)" feature plays a key role in the level of deprivation. Despite this important common feature, Trieste and Cagliari display different characteristics in their principal components.

For Trieste, the first three principal components (PCs) are analysed (Figure 2), as they together explain 74.5% of the total variance. The first PC is determined by the SED-I, the households' size feature (one-person households as a % out of total households) and by the share of individuals in households benefiting from social exclusion transfers (as a percentage of the total population). The second PC is defined by the age composition of the population: the variable representing the highest loading is the share of elderly population (65+ years old) as a % of the total population. The third PC is mainly explained by the employment rate (15+ years old) and by the sex composition of the population (M/F ratio).

Figure 2– *PCA results for Trieste (Istat, 2021).*



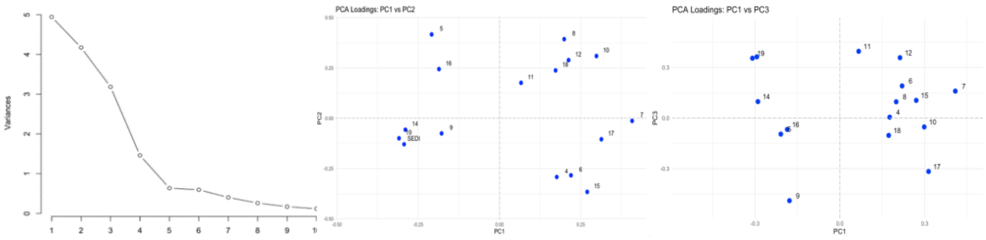
A. PCA scree plot

B. PCA loadings axis 1-2

C. PCA loadings axis 1-3

For Cagliari, the first three principal components (PCs) are presented as well since they together explain 76.8% of the total variance (Figure 3). The first PC is determined by the employment rate (15+ years old) and by university graduates aged 25-64 (% of the population aged 25-64), therefore both variables refer to the socio-economic condition. The second PC is defined by household and age compositions: size of households (population/households) and young people (0-24 years old) as a % of the total population. The third PC is characterized by sex and age compositions: M/F ratio and elderly (65+ years old) as a % of the total population.

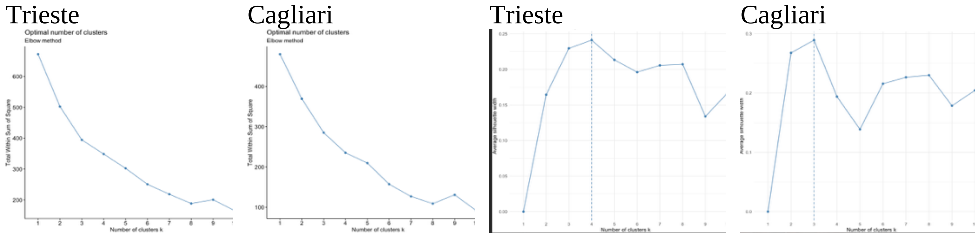
Figure 3 – *PCA results for Cagliari (Istat, 2021).*



A. PCA scree plot *B. PCA loadings axis 1-2* *C. PCA loadings axis 1-3*

These differences in PCA structures between Trieste and Cagliari reflect structural socio-economic divergences in the two municipalities, as described in Section 2. In this regard, the following K-Means cluster analysis shows similarities and differences among SMAs, separately for Trieste e Cagliari. For both municipalities, the number of optimal clusters have been selected equal to three using the Silhouette and the Elbow Method (James *et al.*, 2021) as shown in Figure 4.

Figure 4 - *Elbow and Silhouette methods for Trieste and Cagliari.*

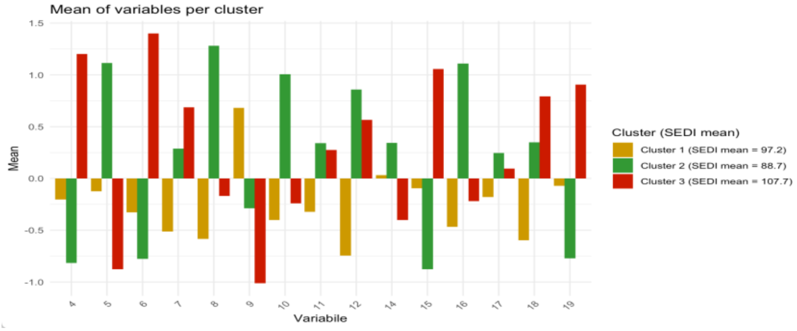


Trieste Cagliari Trieste Cagliari

For both municipalities the SED-I and the set of contextual demographic variables have been included in the computation, after being standardized. The results of the K-Means cluster analysis for the municipality of Trieste have identified cluster number 3 as the most vulnerable cluster as it is characterised by districts with a higher average value of the SED-I.

Figure 5 shows the averages values for each variable in the different clusters. The variables are standardized to make meaningful comparison of the differences between cluster 1, 2 and 3 (the most vulnerable one). The latter is characterized by higher values of population density (population/area), foreigners (% as a share of the total population), one-person households and individuals in households benefiting from social exclusion transfers; on the other hand, is also characterised by a lower value of elderly (65+ years old) as a % of the total population.

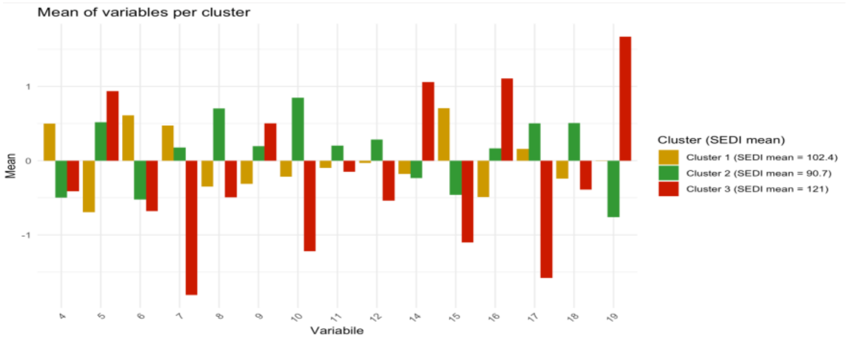
Figure 5 – *K-Means cluster analysis of Trieste (Istat, 2021).*



The results of the K-Means cluster analysis for the municipality of Cagliari have also identified cluster number 3 as the most vulnerable cluster as it is characterised by SMAs with a higher average value of the SED-I. In fact, the characterisation of this cluster is different, compared to the first and the second cluster (Figure 6), because it is characterised by lower values of employment rate (15+ years old), student % as a share of population, university graduates aged 25-64 (% of the population aged 25-64), one-person households and foreigners (% as a share of the total population).

On the other hand, the third and more deprived cluster is characterized by higher values of households with 5 or more members (% out of total households) and foreigners 65+ years old (% as a share of the total foreigners). Finally, this cluster, like what observed for Trieste, has a higher value of individuals in households benefiting from social exclusion transfers (% as a share of the total population).

Figure 6 – *K-means cluster analysis of Cagliari (Istat, 2021).*



Figures 7 and 8 show the maps of the two municipalities under analysis, in which the SMA are categorised on the left with the SED-I and on the right with the assignment to the clusters we obtained in our analysis. We notice that the classification of socio-economic deprivation by SMA is quite similar with both methods, except for one SMA of Cagliari ("Mulinu Becciu") which has very different socio-demographic characteristics from the others. This SMA is in fact characterized by the higher share of elderly people (65+ years old), lower employment and education levels, but it also shows good economic conditions.

Figure 7 – *Comparison of the level of deprivation of Trieste (Istat, 2021).*

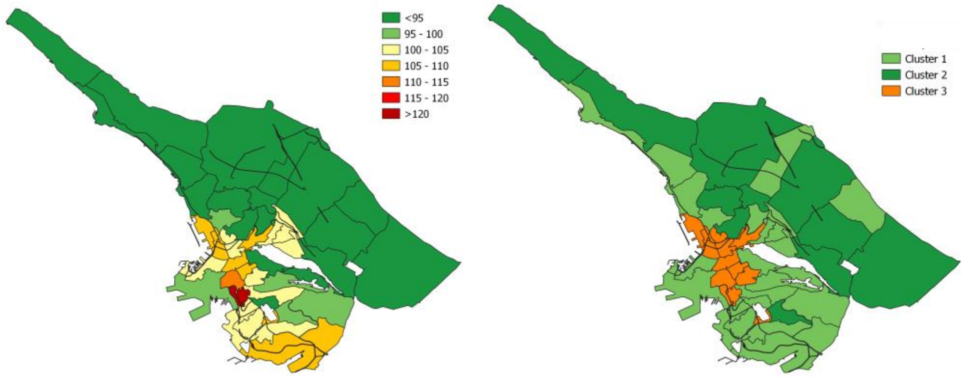
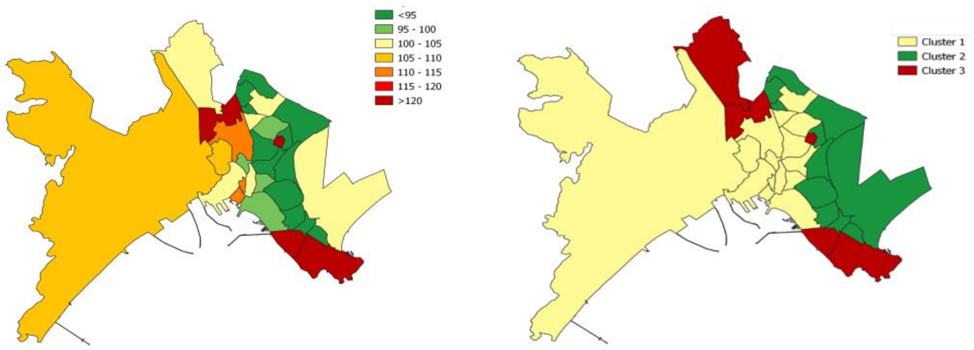


Figure 8 – *Comparison of the level of deprivation of Cagliari (Istat, 2021).*



This paper presents the first attempt to classify different sub-municipal areas with the same methodology. However, some previous information is available for Cagliari (Istat, 2024), and our results align and extend them.

5. Conclusions

This analysis offers a detailed characterization of deprived areas across the two municipalities of Trieste and Cagliari, highlighting distinct socio-demographic patterns. In Trieste, deprivation – as measured by the Social and Economic Deprivation Index (SED-I) – is primarily social in nature. It is closely linked to a significant share of people living alone and the presence of foreign populations. In contrast, deprivation in Cagliari is more strongly associated with economic hardship and lower educational attainment, indicating structural challenges related to employment and human capital development. These differences underscore the importance of tailoring local interventions to the specific profiles of vulnerability within each area.

The analysis was enhanced through the application of two machine learning methodologies: Principal Component analysis and K-means cluster analysis. The first enabled the identification of the most influential variables associated with the SED-I, highlighting the central role of social exclusion transfers, education, and demographic characteristics in shaping vulnerability. K-means clustering allowed us to group municipalities into internally homogeneous clusters, uncovering latent patterns in the data and enabling a differentiated interpretation of local vulnerability dynamics. The findings are robust across different methodologies and geographic contexts and offer useful evidence to support targeted policy interventions. In particular, the consistent association between high deprivation and social exclusion transfers, as well as the protective effects of employment, education, and youth presence, provide actionable insights for designing effective, place-based social and economic strategies.

References

ALLIK M., BROWN D., DUNDAS R., LEYLAND A. H. 2016. Developing a new small area measure of deprivation using 2001 and 2011 census data from Scotland. *Health & Place*, Vol. 39, pp. 122–130.

CARBONETTI G., BIASCIUCCI F., CUTILLO A., MAZZIOTTA M., QUONDAMSTEFANO V., TAMBURRANO M.T., TRONU D. 2025a. Measuring socio-economic deprivation at sub-municipal level through the integration of census and administrative data. *Italian Journal of Economic, Demographic and Statistical Studies – RIEDS*, Vol. LXXIX No.1 Gennaio-Marzo 2025.

CARBONETTI G., BIASCIUCCI F., CUTILLO A., MAZZIOTTA M., QUONDAMSTEFANO V., TRONU D. 2025b. An innovative approach for the analysis of socio-economic phenomena at sub-municipal level: the household deprivation study project. Data, Statistics and AI for Well-Being of People and Organizations. *Book of Short Papers of the ASA Rome Conference*, pp. 121-126. Padua: Cleup. https://doi.org/10.26398/asaproc.0090

HARTIGAN J. A., WONG M. A. 1979. A K-Means Clustering Algorithm, *Journal of the Royal Statistical Society Series C: Applied Statistics*, Vol. 28, Issue 1, March 1979, pp. 100–108.

ISTITUTO NAZIONALE DI STATISTICA 2024. *Sicurezza e stato di degrado delle città e delle loro periferie, Audizione presso la Commissione parlamentare di inchiesta sulle condizioni di sicurezza e sullo stato di degrado delle città e delle loro periferie*, 26 giugno 2024, Istat, Roma, 2024.

ISTITUTO NAZIONALE DI STATISTICA 2020. *Le misure della vulnerabilità: un'applicazione a diversi ambiti territoriali*, Istat, Roma, 2020.

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R. 2021. Unsupervised Learning. In: *An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer, New York, NY.

MAZZIOTTA M., PARETO A. 2017. Synthesis of indicators: the composite indicators approach. In Maggino F. (Ed.) Complexity in Society: From Indicators Construction to their Synthesis. *Social Indicators Research Series*., Springer, pp.159-191.

MAZZIOTTA M., PARETO A. 2016. On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phenomena. *Social Indicators Research*, Vol. 127, No. 3, pp. 983-1003.

Matilde BONELLI, University of Bologna, matilde.bonelli@studio.unibo.it
Giancarlo CARBONETTI, Istat, carbonet@istat.it
Elena GRIMACCIA, Istat, elgrimac@istat.it
Debora TRONU, Istat, tronu@istat.it