symmetry

MDPI

Article

A Novel Two-Level Entropy-Weighted Fuzzy C-Means Algorithm and Its Application for Classifying Urban Patterns by Residential Building Characteristics

Rosa Cafaro [1], Barbara Cardone [1] and Ferdinando Di Martino [1,2,*]

[1] Department of Architecture, University of Naples Federico II, Via Toledo 402, 80134 Napoli, Italy; rosa.cafaro@unina.it (R.C.); b.cardone@unina.it (B.C.)
[2] Center for Interdepartmental Research "Alberto Calza Bini", University of Naples Federico II, Via Toledo 402, 80134 Napoli, Italy
* Correspondence: fdimarti@unina.it; Tel.: +39-081-2538904

Abstract

In this work, a novel entropy-weighted fuzzy c-means variation, referred to as Group-based Entropy Weighted Fuzzy C-Means (GEWFCM), is proposed. This variation introduces a semantic level of partitioning of features into groups. This approach enables the provision of optimal semantic meaning to the clusters, thereby capturing the intrinsic structure of the features, which are naturally grouped into homogeneous semantic sets; the weights are independent of the clusters. The cluster weights provide a direct measure of the importance of each group, determining which dimensions of the phenomenon are relevant, and the intragroup weights determine the most relevant features within a group. Additionally, GEWFCM is computationally more efficient than other cluster-specific weighted fuzzy clustering algorithms, due to the independence of the weights from the clusters. The efficacy of the method was assessed by evaluating census data from 16 Italian cities, with the objective of partitioning urban settlements based on characteristics of residential buildings, including construction technique, period, number of floors, and state of conservation. The findings suggest that the proposed algorithm effectively captures the semantic meaning of clusters. In addition, a comparative analysis between GEWFCM and the well-known Entropy Weighted Fuzzy C-Means (EWFCM) algorithm showed that, although both algorithms provide high similarity of results for all case studies, GEWFCM is significantly faster.

Keywords: EWFCM; GEWFCM; weighted FCM; residential buildings

Academic Editor: Jie Yang

Received: 14 April 2026
Revised: 30 April 2026
Accepted: 7 May 2026
Published: 8 May 2026

1. Introduction

The classic Fuzzy C-Means (FCM) algorithm [1,2] operates under the implicit assumption that all features contribute equally to the definition of clusters. This assumption is frequently deemed unrealistic in real-world datasets, which are often characterized by heterogeneous, redundant, or irrelevant features. In order to surmount this limitation, a number of studies have proposed variants of FCM based on the introduction of features associated with weights. The objective of this approach is to adapt the distance metric to the data structure.

In recent years, several weighted variations of FCM have been proposed with the objective of balancing the importance of features in cluster construction. A weighted FCM feature selection method based on the principle of refined justifiable granularity, measuring the significance of features in the feature space, is proposed in [3]. In [4], a classification method is proposed that integrates a weighted FCM algorithm and an enhanced adaptive neuro-fuzzy inference model for the classification of chronic kidney disease. In [5], an algorithm based on a dissimilarity measure is employed for the purpose of clustering gene expression data.

A significant constraint of these methodologies is the complexity involved in assigning a semantic interpretation to the clusters. To address this limitation, Keller and Klawonn [6] propose a model in which each feature possesses a cluster-dependent weight, that is, a distinct parameter for each feature–cluster pair. In this formulation, each cluster is characterized by its own relevant subspaces. Furthermore, features may possess differing levels of importance across different clusters. This approach exhibits a greater degree of expressiveness in comparison with FCM; however, it concomitantly introduces novel challenges, including a substantial augmentation in the number of parameters, a heightened computational complexity, and an augmented sensitivity to noise and initialization.

In [7], a weight learning mechanism based on optimization techniques (e.g., gradient descent) is proposed, showing that an appropriate weight assignment can improve clustering quality compared to the standard FCM. This algorithm has been demonstrated to exhibit superior speed and resilience to noise compared to [6]. However, it should be noted that these weights are global, which means they do not take into account the differences between clusters. Additionally, the estimation of these weights can be unstable and dependent on the initialization process.

In order to enhance the interpretability of the clusters, an entropy-weighted FCM variation based on entropy regularization, termed Entropy-Weighted Fuzzy C-Means (EWFCM), is proposed in [8]. This approach involves differentiating the weight of the features between the clusters. In EWFCM, feature weights are determined on a per-cluster basis and undergo an exponential transformation of feature costs; an entropy term is incorporated to circumvent degenerate solutions. EWFCM demonstrates superior robustness in comparison with non-regularized methods and exhibits enhanced adaptability in the context of high-dimensional data. Nonetheless, the model exhibits considerable limitations. Primarily, it possesses a high degree of computational complexity. Indeed, the number of model parameters is proportional to the product of the number of features and the number of clusters. This results in a high cost per iteration, protracted convergence times, and suboptimal scalability when dealing with large datasets.

In order to manage high-dimensional datasets, a feature-weighted entropy FCM method was proposed in [9]. This method allows for the reduction of the feature space by removing irrelevant features with small weights. This method automatically calculates individual feature weights while reducing redundant feature components, thereby enabling the clustering of high-dimensional features.

In [10], a variation of EWFCM is proposed. In this variation, a subset of the feature space is extracted, and a weight is allocated to each feature dimension based on the feature's impact on clustering. In [11], an automatic local feature weighting and cluster weighting mechanism is proposed to properly weigh the features and to attenuate the initialization sensitivity of FCM.

An adaptive feature-weighted entropy FCM algorithm for image segmentation is applied in [12] to mitigate the contribution of less significant features. The authors employ a distance metric that incorporates both Euclidean and non-Euclidean distances.

In [13], a novel weighted FCM is proposed, in which an objective function is utilized that is based on a feature-weighted generalized entropy regularization strategy.

These approaches, while endeavoring to reduce the computational complexity of EWFCM, fail to fully capture the semantics of the data and efficiently optimize the interpretability of the results. Additionally, their computational complexity remains high, rendering them ill-suited for the management of high-dimensional datasets.

Table 1 summarize the characteristics, strengths, and limitations of the existing method types. In particular, while EWFCM offers high flexibility thanks to cluster-specific weighting, it suffers from increased computational complexity and limited interpretability.

Table 1. Characteristics, strengths, and limitations of the existing weighted FCM methods.

Type of Method	Core Idea	Complexity	Interpretability	Strengths	Limitations
FCM ([1,2])	Prototype-based clustering	Low	Low	Simple, fast	No feature weighting; sensitive to noise
Grouped FCM ([6])	Feature grouping	Low	Low	Structured representation	Limited adaptability; sensitive to noise
wFCM ([7])	Global feature weighting	Low	Low	Reduces noise	Does not exploit semantic relationships between features; all variables are treated independently
EWFCM-based ([8–13])	Entropy-based weighting	High	Medium-Low	Adaptive feature selection	High computational cost; low interpretability;

In a multitude of real-world application contexts, datasets manifest an inherent structural organization, wherein variables can be naturally classified into homogeneous semantic groups. These sets reflect well-defined conceptual categories and not simple arbitrary aggregations of features.

For instance, the criteria employed to assess and appraise the quality of drinking water can be categorized into several domains. These domains include physical characteristics, such as temperature, turbidity, and color; chemical characteristics, such as pH, fixed residue, and hardness; chemical–toxicological characteristics, such as the concentration of heavy metals and organic pollutants; and microbiological characteristics, such as bacterial concentrations and indicators of environmental contamination.

The grouped FCM algorithm [6] while adapting to this natural way of grouping features into categories, does not capture the different influences of the features in a group, since all the features within a group are treated in the same way, which reduces the model's ability to capture heterogeneous feature relevance.

In this scenario, traditional fuzzy clustering methods and their feature-weighted extensions, such as EWFCM, treat variables independently, neglecting the hierarchical or semantic structure of the data.

In order to overcome this limitation, a variation of EWFCM was proposed. This variation is referred to as Group-based EWFCM (GEWFCM). GEWFCM introduces a two-level weighting mechanism. In this mechanism, feature groups represent higher-level semantic units. Furthermore, features within each group contribute relatively to the representation of clusters.

In the proposed model, the explicit introduction of groups enables the objective function to be modeled consistently with the data structure. Specifically, the weight assigned to a feature is expressed as the product of the weight assigned to the group to which the feature belongs and the weight of the feature within that group.

In comparison to EWFCM, in which weights are defined independently for each cluster and each feature, the proposed model introduces a structural regularization that reduces uncontrolled weight variability and improves solution stability.

Unlike existing feature-weighted approaches, the proposed GEWFCM does not simply reduce the number of parameters but introduces a structurally constrained weighting model that explicitly reflects the semantic organization of the feature space. In particular, while EWFCM assigns feature weights independently for each cluster, GEWFCM decomposes feature importance into two complementary components: a group-level weight and an intra-group feature weight. This formulation allows the model to capture both the relevance of semantic dimensions and the contribution of individual features within each dimension.

This represents a conceptual shift from unstructured feature weighting to semantically guided clustering. As a result, GEWFCM not only reduces the dimensionality of the optimization problem, but also improves the stability of the solution and provides a more interpretable representation of clusters.

Furthermore, compared to existing grouped approaches, GEWFCM overcomes the limitation of uniform feature importance within groups by introducing intra-group weighting, thereby enabling a more flexible and accurate modelling of heterogeneous feature relevance.

A further objective of the method is to improve the interpretability of the results. In the context of EWFCM, the weights are designated as cluster-specific, not constrained by a semantic structure. This implies that each cluster is characterized by combinations of features that are challenging to interpret and may not be consistent with each other.

In contrast, in GEWFCM, the weights are cluster independent. The group weights provide a direct measure of the importance of each semantic dimension, determining which dimensions of the phenomenon are relevant. The intra-group weights allow us to identify the most relevant features within each group and determine which specific variables contribute to the definition of the clusters. This dual structure is intended to ensure greater semantic interpretability of the clusters.

Additionally, the independence of the weights from the clusters results in enhanced computational efficiency in comparison to EWFCM. This results in a significant reduction in the dimensionality of the optimization problem, greater numerical stability, and shorter convergence times, making GEWFCM suitable for handling high-dimensional data.

In summary, the proposed method is characterized by two main contributions:

- Improved interpretability: Partitioning features into groups allows for better semantic meaning to be assigned to clusters. This dual structure enables a two-level interpretation of clusters: a global level for groups, which allows us to determine which dimensions of the phenomenon are relevant, and a local level consisting of features within a group, which allows us to determine which group-specific variables contribute to the definition of a cluster.
- Higher computational efficiency: Indeed, the number of parameters is reduced compared to cluster-specific models like EWFCM. The independence of the weights from clusters reduces computational cost, especially when dealing with many clusters.

GEWFCM has been tested as an unsupervised classifier for classifying urban settlements based on a set of residential building characteristics acquired from the population and building census dataset compiled by the Italian National Statistical Institute (ISTAT). The information on residential buildings is summarized by census zone; for each characteristic, the number of residential buildings with that characteristic located in each census zone is measured. These characteristics are grouped into five types: construction technique, construction period, number of floors, number of interiors, and state of conservation.

https://doi.org/10.3390/sym18050807

After introducing the EWFCM algorithm in Section 2, Section 3 discusses the proposed algorithm in detail and describes the case studies used. Section 4 presents the test results and discusses the comparative results. Section 5 includes concluding remarks.

2. Preliminary Concepts

2.1. The EWFCM Algorithm

EWFCM is an extension of FCM that, instead of considering each feature of equal importance, assigns them a weight to enhance their contribution in the creation of clusters.

Let $x_i \in \mathbb{R}^p \;\; i = 1, \ldots, N$ be the set of N samples Let X be a set of N samples described by p features and let C be the number of clusters.

The fuzzy partition matrix is denoted by U, the membership degree of the ith sample in the kth cluster by u_{ik}, and the centroid of kth cluster by v_k.

In EWFCM, the relevance of the jth feature in the kth cluster is represented by a cluster-dependent feature weight w_{kj} where $0 \leq w_{kj} \leq 1$ and $\sum_{k=1}^{C} w_{kj} = 1$.

The objective function is given by

$$J(U, V, w) = \sum_{i=1}^{N} \sum_{k=1}^{C} {u_{ik}}^{m} \sum_{j=1}^{p} w_{jk} \left(x_{ij} - v_{kj}\right)^2 + \lambda \sum_{k=1}^{C} \sum_{j=1}^{p} w_{jk} \log\left(w_{jk}\right) \quad (1)$$

where m is the fuzzifier parameter and Γ is an entropy parameter set for the regularization of the weights.

The first term is exactly the classical fuzzy energy as in FCM, but with the distance scaled by the eights w_{kj}. The last term is an entropy penalty term that prevents degenerate solutions (e.g., all the weight on a single feature). This term encourages the membership degrees to be more distributed, to avoid the model getting stuck on local optima.

Using the Lagrange multiplicators to minimize the objective function, the three solutions for the partition matrix, the centroids of the clusters and the weights are obtained.

$$v_{kj} = \frac{\sum_{i=1}^{N} u_{ki}^{m} x_{ij}}{\sum_{i=1}^{N} u_{ki}^{m}} \quad (2)$$

$$u_{ki} = \left(\sum_{r=1}^{C} \left(\frac{\sum_{j=1}^{p} w_{kj} \left(x_{ij} - v_{kj}\right)^2}{\sum_{j=1}^{p} w_{rj} \left(x_{ij} - v_{rj}\right)^2} \right)^{\frac{1}{m-1}} \right)^{-1} \quad (3)$$

$$w_{kj} = \frac{\exp\left(- \frac{\sum_{i=1}^{N} u_{ki}^{m} \left(x_{ij} - v_{kj}\right)^2}{\Gamma} \right)}{\sum_{s=1}^{p} \exp\left(- \frac{\sum_{i=1}^{N} u_{ki}^{m} \left(x_{is} - v_{ks}\right)^2}{\Gamma} \right)} \quad (4)$$

In (4) the weight are given in exponential form as softmax functions [14,15].

The process is iterated until a stop criterion is reached.

Below is shown in pseudocode the Algorithm 1 EWFCM.

Algorithm 1: EWFCM

Input: Set of data points
Number of cluster C
Fuzzifier parameter m

Entropy parameter Γ
Stop iteration error ε
Output: Partition matrix
Centroids of the clusters
Weights

Set initially randomly the centroids of the clusters and the weights
Repeat
  Update the partition matrix by (2)
  Update the centroids by (3)
  Update the weights by (4)
Until $(\left|J\left(U^{(t+1)}, V^{(t+1)}, w^{(t+1)}\right) - J\left(U^{(t)}, V^{(t)}, w^{(t)}\right)\right| > \varepsilon)$ **and** (numIter ≤ maxIter)
Return $U^{(t+1)}, V^{(t+1)}, w^{(t+1)}$

3. Materials and Methods

3.1. The GEWFCM Algorithm

Let $x_i \in \mathbb{R}^p \ \ i = 1, \ldots, N$ be N samples. Let the p features partitioned into H disjoint semantic groups $G_1, \ldots, G_{\mathrm{H}}$, where $G_{h(j)}$ denotes the group to which the *j*th feature belongs.

GEWFCM introduces a two-level weighting mechanism. To each group G_{h} is assigned a weight $g_h$where $g_h \geq 0$ and $\sum_{h=1}^{H} g_h = 1$. To the jth feature belonging to the group $G_{h(j)}$ is assigned an intra-group weight $w_{j|h}$ where $w_{j|h} \geq 0$ and $\sum_{j \in G_h} w_{j|h} = 1$.

The aggregate weight associated with the *j*th feature is therefore defined as follows:

$$w_j \ = \ g_{h(j)} \ w_{j|h(j)} \tag{5}$$

where the following constraint holds:

$$\sum_{j=1}^{p} w_j = 1 \tag{6}$$

The objective function is given by

$$J(U, V, g, w) = \sum_{i=1}^{N} \sum_{k=1}^{C} u_{ik}^{m} \left(\sum_{j=1}^{p} w_j \, (x_{ij} - v_{kj})^2 \right) + \Gamma_g \sum_{h=1}^{H} g_h \log \ g_h + \sum_{h=1}^{H} \Gamma_h \sum_{j \in G_h} w_{j|h} \log \ w_{j|h} \tag{7}$$

Here, $U = (u_{ik})$ is the fuzzy partition matrix, $V = (v_k)$ is the set of cluster centroids, and $m > 1$ is the fuzzifier.

The parameter $\Gamma_g > 0$ controls the entropy regularization of the group weights, whereas $\Gamma_h > 0$ controls the entropy regularization of the intra-group feature weights. Small values of γ_g and γ_h promote concentrated weight distributions; large values produce smoother and more uniform distributions.

The convergence properties of GEWFCM are consistent with those of standard FCM and its weighted extensions. GEWFCM follows an alternating optimization scheme, where the objective function is minimized with respect to one set of variables at a time while keeping the others fixed.

Specifically, the algorithm alternates between the update of membership degrees, cluster centroids, intra-group feature weights, and group weights. Each of these steps corresponds to the solution of a constrained optimization problem, which ensures that the value of the objective function does not increase. Therefore, the sequence of objective

function values $J^{(t)}$ at the iteration t is monotonically non-increasing and $J^{(t+1)}) \leq J^{(t)}$. Moreover, since the objective function is bounded from below, the sequence converges to a stationary point. Although global optimality cannot be guaranteed due to the non-convex nature of the problem, this property ensures the stability of the algorithm.

In particular, GEWFCM minimizes the objective function by an alternating optimization procedure. At iteration t, the aggregate weights are computed from (5) and normalized so that their sum is equal to one. The data are transformed according to the aggregate feature weights:

$$\tilde{x}_{ij}^{(t)} = \sqrt{w_j^{(t)}}\, x_{ij}. \tag{8}$$

Standard FCM is applied to the transformed data, yielding the updated memberships

$$u_{ik}^{(t+1)} = \left(\sum_{r=1}^{C} \left(\frac{d_{ik}^{(t)}}{d_{ir}^{(t)}} \right)^{\frac{1}{m-1}} \right)^{-1} \tag{9}$$

where $d_{ik}^{(t)} = \| \tilde{x}_i^{(t)} - \tilde{v}_k^{(t)} \|^2$, and the centroids in the transformed space are

$$\tilde{v}_k^{(t+1)} = \frac{\sum_{i=1}^{N} \left(u_{ik}^{(t+1)}\right)^m \tilde{x}_i^{(t)}}{\sum_{i=1}^{N} \left(u_{ik}^{(t+1)}\right)^m} \tag{10}$$

The centroids in the original feature space are then updated as

$$v_{kj}^{(t+1)} = \frac{\sum_{i=1}^{N} \left(u_{ik}^{(t+1)}\right)^m x_{ij}}{\sum_{i=1}^{N} \left(u_{ik}^{(t+1)}\right)^m}. \tag{11}$$

For fixed U and V, the intra-group weights are obtained by minimizing the part of the objective function that depends on the weights of the features belonging to the same group. The feature cost is defined as

$$S_j = \sum_{i=1}^{N} \sum_{k=1}^{C} u_{ik}^m \left(x_{ij} - v_{kj}\right)^2. \tag{12}$$

The quantity S_j measures the intra-cluster dispersion associated with the j-th feature. Lower values of S_j indicate that the feature is more effective in describing the cluster structure.

For each group G_h, the intra-group weights are obtained by minimizing the weight-dependent part of the objective function under the normalization constraint.

$$\min_{\{w_{j|h}\}_{j \in G_h}} \sum_{j \in G_h} w_{j|h}\, S_j + \Gamma_h \sum_{j \in G_h} w_{j|h} \log\left(w_{j|h}\right) \tag{13}$$

subject to $\sum_{j \in G_h} w_{j|h} = 1$.

Introducing the Lagrange multiplier λ_h, the Lagrangian is

$$\mathcal{L}_h = \sum_{j \in G_h} w_{j|h}\, S_j + \Gamma_h \sum_{j \in G_h} w_{j|h} \log\left(w_{j|h}\right) + \lambda_h \left(\sum_{j \in G_h} w_{j|h} - 1 \right). \tag{14}$$

The stationarity condition

$$\frac{\partial \mathcal{L}_h}{\partial w_{j|h}} = S_j + \Gamma_h (1 + \log\left(w_{j|h}\right)) + \lambda_h = 0 \tag{15}$$

leads to the closed-form update

$$w_{j|h}^{(t+1)} = \frac{\exp\left(\frac{S_j^{(t+1)}}{\Gamma_h}\right)}{\sum_{r\in G_h}\exp\left(-\frac{S_r^{(t+1)}}{\Gamma_h}\right)}. \tag{16}$$

The group cost of the hth group G_h is defined as the weighted average feature cost within the group:

$$S_h = \frac{1}{|G_h|}\sum_{j\in G_h} w_{j|h}\, S_j. \tag{17}$$

The quantity S_h measures the average intra-cluster dispersion of the features in group G_h, weighted by their intra-group importance. Lower values of S_h indicate that the corresponding semantic group provides a more informative description of the cluster structure.

For fixed U, V, and $w_{j|h}$, the group weights are obtained by solving

$$\min_{\{g_h\}_{h=1}^{H}} \sum_{h=1}^{H} g_h S_h + \Gamma_g \sum_{h=1}^{H} g_h \log(g_h) \tag{18}$$

subject to $\sum_{h=1}^{H} g_h = 1$.

Introducing the Lagrange multiplier μ, the Lagrangian is

$$\mathcal{L}_g = \sum_{h=1}^{H} g_h S_h + \Gamma_g \sum_{h=1}^{H} g_h \log(g_h) + \mu\left(\sum_{h=1}^{H} g_h - 1\right). \tag{19}$$

The stationarity condition

$$\frac{\partial \mathcal{L}_g}{\partial g_h} = S_h + \Gamma_g(1 + \log(g_h)) + \mu = 0 \tag{20}$$

yields the closed-form update

$$g_h^{(t+1)} = \frac{\exp\left(-\frac{S_g^{(t+1)}(h)}{\Gamma_g}\right)}{\sum_{r=1}^{H}\exp\left(-\frac{S_g^{(t+1)}(r)}{\Gamma_g}\right)}. \tag{21}$$

Finally, convergence is checked through the variation of the aggregate feature weights:

$$\Delta^{(t+1)} = \sum_{j=1}^{p}\left|w_j^{(t+1)} - w_j^{(t)}\right|. \tag{22}$$

If $\Delta^{(t+1)} < \varepsilon$, the iterative process stops; otherwise, the next iteration is performed.

Below, Algorithm 2 summarizes the GEWFCM procedure.

Algorithm 2: GEWFCM

Input:	Set of data points
	Number of cluster C
	Fuzzifier parameter m
	Entropy parameters Γ_g and Γ_h
	Stop iteration error ε
Output:	Partition matrix
	Centroids of the clusters
	Group weights g_h

https://doi.org/10.3390/sym18050807

Intra-group weights $w_{j|h}$

Initialize $U^{(0)}$ (or, equivalently, the centroids), the group weights $g^{(0)}$, and the intra-group weights $w_{j|h}^{(0)}$.

Repeat

Compute the aggregate feature weights $w_j^{(t)}$ from (5) and normalize them
Transform the data according to (8).
Update the partition matrix by (9)
Update the transformed centroids by (10) and the original centroids by (11).
Update the original centroids by (11)
Compute the feature costs by (12)
Update the intra-group weights by (16)
Compute the group costs by (17)
Compute the weight variation $\Delta^{(t+1)}$ by (22)

Until $\Delta^{(t+1)} > \varepsilon$

Return $U^{(t+1)}, V^{(t+1)}, g^{(t+1)}, w^{(t+1)}$

From a computational standpoint, GEWFCM has a significant advantage over EWFCM. In EWFCM, the number of feature weights updated at each iteration is equal to $p \times C$, where p is the number of features and C is the number of clusters. In contrast, GEWFCM requires the update of $p + H$ weights per iteration, where H is the number of semantic groups and typically $H \ll p$.

Consequently, GEWFCM reduces the dimensionality of the optimization problem and generally leads to lower computational cost and faster convergence.

3.2. The Case Studies

GWEFCM has been tested to classify urban settlement zones based on residential building fabric characteristics. To this end, the population and building census datasets carried out by the Italian Institute of Statistics (ISTAT) in 2011 were utilized.

The objective of the present study was to conduct a series of tests. To this end, all information relating to the characteristics of residential buildings was extracted.

The information was grouped by census zone, and the samples included 16 Italian cities. Each piece of information corresponds to the number of residential buildings in the census zone that possess a specific characteristic.

The data have been standardized by dividing it by the total number of buildings in the census zone. Consequently, each feature will contain the frequency of residential buildings exhibiting a specific characteristic.

This normalization corresponds to a frequency-based scaling, which is particularly appropriate for this type of data, where each feature represents the relative frequency of a specific building characteristic within a census zone.

Alternative normalization techniques, such as min–max scaling or z-score standardization, could also be considered. However, these approaches may alter the semantic meaning of the features. In particular, z-score normalization assumes a Gaussian distribution and may introduce negative values, which are not meaningful for frequency-based variables. Min–max normalization, on the other hand, may reduce variability in the presence of outliers.

The adopted normalization preserves the relative proportions of the features within each census zone, which is essential for maintaining the interpretability of the clustering results. Preliminary tests showed that the proposed method is robust with respect to the choice of normalization, and the overall clustering structure remains stable. For these reasons, frequency-based normalization was selected as the most appropriate preprocessing step for this study.

The features were grouped into five groups, as specified in Table 2.

Table 2. Features related to the frequency of residential buildings with a specific characteristic and their grouping.

Group	Feature	Description
Construction technique	d_5	Frequency of residential buildings constructed of masonry
	d_6	Frequency of residential buildings constructed of reinforced concrete
	d_7	Frequency of residential buildings constructed of other materials
Construction period	d_8	Frequency of residential buildings constructed before 1919
	d_9	Frequency of residential buildings constructed from 1919 to 1945
	d_{10}	Frequency of residential buildings constructed from 1946 to 1960
	d_{11}	Frequency of residential buildings constructed from 1961 to 1970
	d_{12}	Frequency of residential buildings constructed from 1971 to 1980
	d_{13}	Frequency of residential buildings constructed from 1981 to 1990
	d_{14}	Frequency of residential buildings constructed from 1991 to 2000
	d_{15}	Frequency of residential buildings constructed from 2001 to 2005
	d_{16}	Frequency of residential buildings constructed after 2005
Number of floors	d_{17}	Frequency of single-story residential buildings
	d_{18}	Frequency of two-story residential buildings
	d_{19}	Frequency of three-story residential buildings
	d_{20}	Frequency of residential buildings with four or more floors
Number of interiors	d_{21}	Frequency of single-family residential buildings
	d_{22}	Frequency of two-apartment residential buildings
	d_{23}	Frequency of residential buildings from three to four apartments
	d_{24}	Frequency of residential buildings from five to eight apartments
	d_{25}	Frequency of residential buildings from nine to sixteen apartments
	d_{26}	Frequency of residential buildings with at least sixteen apartments
State of conservation	d_{28}	Frequency of residential buildings with excellent state of conservation
	d_{29}	Frequency of residential buildings with fair state of conservation
	d_{30}	Frequency of residential buildings with poor state of conservation
	d_{31}	Frequency of residential buildings with very poor state of conservation

In recent years, urban data analysis has increasingly leveraged advanced machine learning frameworks, including deep learning and spatial analytics techniques, for tasks such as urban morphology analysis and Transit-Oriented Development (TOD). These approaches enable the extraction of complex patterns from large-scale urban datasets.

For example, recent studies have explored the integration of clustering and representation learning methods to capture spatial and functional characteristics of urban environments [16]. While these approaches are particularly effective for predictive modeling and large-scale pattern recognition, they often rely on complex architectures and may lack interpretability.

In this context, the proposed GEWFCM method provides a complementary approach, focusing on interpretable unsupervised clustering. By introducing a structured feature weighting mechanism based on semantic grouping, GEWFCM enables a meaningful description of urban patterns while maintaining computational efficiency.

In [17] an unsupervised FCM-based classifier was tested to classify urban patterns based on residential building characteristics related to construction techniques and construction macro-periods.

In the experimental tests conducted on GWEFCM, all the characteristics of the residential buildings present in the datasets provided by ISTAT were considered separately.

For each municipality case study, the dataset was constructed including all the 26 features described in Table 1, where the instances are the census zones into which the municipality is partitioned.

Each census zone will be assigned to the clustering to which it belongs with the highest membership degree.

GEWFCM was implemented using the Python ArcPy libraries from the GIS tool ArcGIS Pro 3.5.

The Xie–Beni validity index [18,19] was employed to ascertain the optimal number of clusters. Xie–Beni determines the optimal number of clusters by minimizing the ratio between the compactness of the clusters (intra-cluster variance) and the minimum separation between the cluster centers. Xie–Beni is the most widely used validity index in FCM to determine the optimal number of clusters [20].

Several samplings were performed to determine the optimal values of the Γ_g and Γ_h parameters. In each trial, different combinations of the two parameters were set, and the combination that generated the minimum value of the Xie–Beni index was selected. The best values are obtained setting $\Gamma_g = 10$ and $\Gamma_h = 10$.

To give semantic meaning to the clusters, the centroid values were normalized by dividing them by the sum of the centroid values of the features in the corresponding group. Then, a linguistic label is assigned as the relevance of each feature in a cluster in the following way.

Let v_{kj} be the value of the jth component of the centroid of the kth cluster.

Let h(j) be the group in which the jth feature is inserted and let $|h(j)|$ the cardinality of this group.

The relevance of the jth feature in the cluster is labeled as *significant* if $v_{kj} > 2/|h(j)|$. Otherwise, it is labeled as *not negligible* if $v_{kj} > 1/|h(j)|$ or *negligible* if $v_{kj} \leq 1/|h(j)|$.

For example, suppose that in a given cluster the normalized values obtained for the Construction technique group features are $d_5 = 0.8$, $d_6 = 0.15$, $d_7 = 0.05$. In this case, $|h(j)| = 3$ and the relevance of each of the three features in the cluster will be, respectively,

r_5 = *significant* r_6 = *negligible* r_7 = *negligible*

This cluster will then group together urban areas with a significant prevalence of masonry buildings.

4. Results and Discussion

The experimental tests were carried out by acquiring ISTAT census data on building characteristics for the 16 most populous Italian cities. The data sources are the ISTAT census datasets conducted in 2011 on all Italian municipalities. They are available at https://www.istat.it/notizia/basi-territoriali-e-variabili-censuarie, accessed on 1 February 2026.

For each city, a dataset was extracted containing data relating to the frequency of residential buildings with the characteristics described in Table 1. For reasons of brevity, the results obtained for the cities of Genoa, Bari, and Naples are shown in detail and the results obtained for all cities are summarized.

In Table 3, the values assigned to the GEWFCM parameters are shown.

Table 3. Values of the parameters used by executing GEWFCM.

Parameter	Description	Value
m	Fuzzifier	2
ε	Stop iteration error	1×10^{-5}
Γ_g	Groups entropy parameter	10
Γ_h	Features entropy parameter	10

As shown in [21], although the optimal choice for the fuzzifier parameter m depends on the dataset, the optimal range is between 1.5 and 2.5, and the central value of m = 2 is considered a safe and robust choice; it is a well-established best practice in literature that ensures good management of uncertainties and overlaps in the data, avoiding both excessive crispness (m tending towards 1) and excessive blurriness (m greater than 2).

The value of the convergence threshold ε was set to 1×10^{-5} because it is small enough to ensure that the cluster centers have stabilized and do not undergo significant changes. A lower threshold would increase the number of iterations required for convergence, without significantly improving the quality of the clustering.

The group and feature entropy parameters Γ_g and Γ_h were selected through a systematic exploration over the range [1, 100]. For each combination of the two parameters, GEWFCM was executed and the corresponding Xie–Beni validity index was computed.

The optimal values were selected as those minimizing the Xie–Beni index. In addition, a sensitivity analysis was conducted to assess the robustness of the proposed method with respect to variations of Γ_g and Γ_h The experimental results show that the clustering structure remains stable over a relatively wide range of parameter values, indicating that the method is not overly sensitive to the specific choice of these parameters. This analysis confirms that the selected values of Γ_g and Γ_h provide a good trade-off between sparsity of the weights and stability of the clustering results.

Inn each test, a preprocessing phase is performed to determine the optimal number of clusters. This is accomplished by running GWEFCM while increasing the number of clusters from C = 2 to C = 10. The optimal number of clusters is obtained for the value of C that minimizes the Xie–Beni index.

We performed 20 independent runs of each algorithm with different random initializations for each test case. The results are reported in terms of average.

All experiments were conducted on a machine equipped with an Intel Core i7-12700K processor and 32 GB of RAM. The GEWFCM and EWFCM algorithms was implemented using the Python 3.13 programming language and the NumPy, SciPy, and scikit-learn libraries. The suite ESRI ArcGIS Pro 10.3 release was used to construct all the thematic maps. Both GEWFCM and EWFCM were implemented in the same environment and executed under identical conditions to ensure a fair comparison.

4.1. Building Classification of Genoa

In the 2011 ISTAT census database Genoa is partitioned in 3616 census zones, of which 3454 are residential (Figure 1).

The input dataset was prepared by selecting only the residential census zones and constructing the 26 features as in Table 1.

At the end of the preprocessing phase, the optimal number of clusters was obtained with value C = 3, for which the smallest value of the Xie–Beni index was measured to be equal to 0.660.

For C = 3 the convergence is reached after eight iterations, in which the difference Δ between the weights obtained in the last iteration and those obtained in the previous iteration (17) is 2.139×10^{-6}.

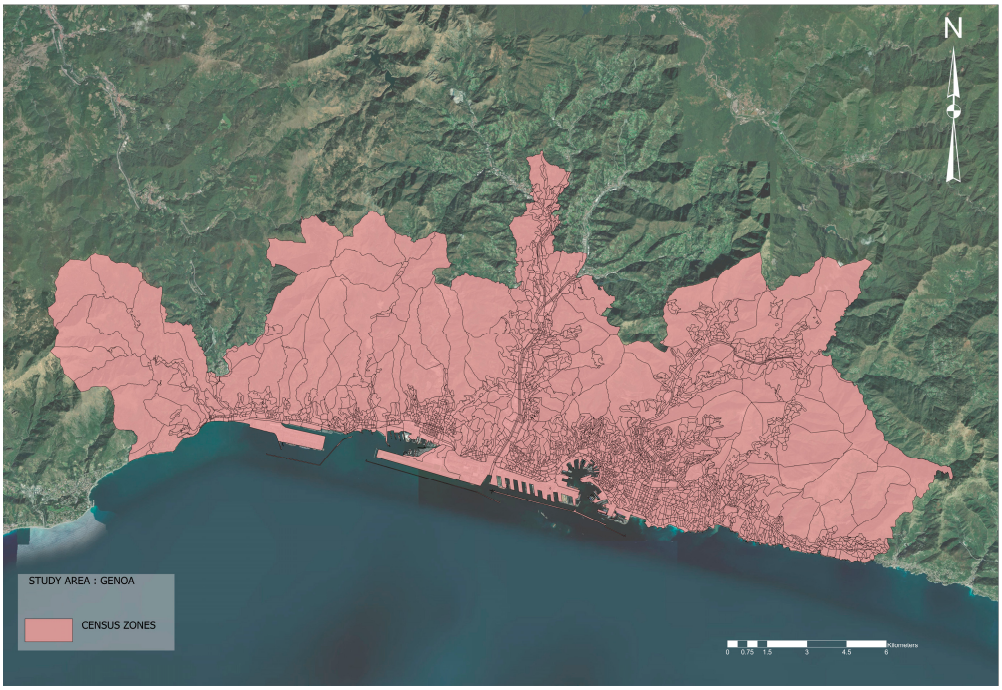

Figure 1. Partitioning of the city of Genoa into census zones.

Then, the census zones of Genoa were grouped into three clusters, C_1, C_2, and C_3. Table 4 shows, normalized with respect to the groups to which they belong, the centroids of the three clusters and the relevance of the features with respect to the three clusters.

Table 4. Cluster centroids and feature relevance for Genoa.

Group	Feature	Cluster Centroids			Relevance		
		C_1	C_2	C_3	r_1	r_2	r_3
Construction technique	d_5	0.762	0.384	0.599	significant	not negligible	not negligible
	d_6	0.188	0.668	0.346	negligible	significant	not negligible
	d_7	0.050	0.018	0.055	negligible	negligible	negligible
Construction period	d_8	0.526	0.200	0.331	significant	not negligible	significant
	d_9	0.213	0.219	0.209	not negligible	not negligible	not negligible
	d_{10}	0.140	0.277	0.286	not negligible	significant	significant
	d_{11}	0.072	0.190	0.093	negligible	not negligible	negligible
	d_{12}	0.028	0.074	0.041	negligible	negligible	negligible
	d_{13}	0.014	0.032	0.027	negligible	negligible	negligible
	d_{14}	0.003	0.005	0.006	negligible	negligible	negligible
	d_{15}	0.002	0.003	0.004	negligible	negligible	negligible
	d_{16}	0.002	0.002	0.004	negligible	negligible	negligible
Number of floors	d_{17}	0.099	0.039	0.087	negligible	negligible	negligible
	d_{18}	0.279	0.100	0.234	not negligible	negligible	negligible
	d_{19}	0.209	0.114	0.200	negligible	negligible	negligible
	d_{20}	0.412	0.747	0.479	not negligible	significant	not negligible
Number of interiors	d_{21}	0.242	0.095	0.227	not negligible	negligible	not negligible
	d_{22}	0.215	0.079	0.165	not negligible	negligible	negligible
	d_{23}	0.119	0.072	0.111	negligible	negligible	negligible
	d_{24}	0.109	0.099	0.105	negligible	negligible	negligible
	d_{25}	0.107	0.158	0.114	negligible	negligible	negligible
	d_{26}	0.208	0.497	0.278	not negligible	significant	not negligible
State of conservation	d_{28}	0.221	0.163	0.256	negligible	negligible	not negligible
	d_{29}	0.611	0.729	0.543	significant	significant	significant

d_{30}	0.151	0.100	0.180	negligible	negligible	negligible
d_{31}	0.016	0.008	0.021	negligible	negligible	negligible

The Cluster C1 category is comprised of census zones that predominantly feature residential buildings constructed in load-bearing masonry prior to 1919 and that have undergone a relatively intact conservation process. Cluster C2 comprises census zones that are distinguished by the predominant use of reinforced concrete in residential building construction during the post-war period, spanning from 1945 to 1960, and are notable for their state of preservation. The third cluster encompasses census zones characterized by the coexistence of residential buildings in good repair, constructed using load-bearing masonry prior to 1919, and those constructed using reinforced concrete between 1945 and 1960.

The thematic map in Figure 2 illustrates the partitioning of the census zones of Genoa into three clusters. Non-residential census zones are indicated by the use of the color gray.

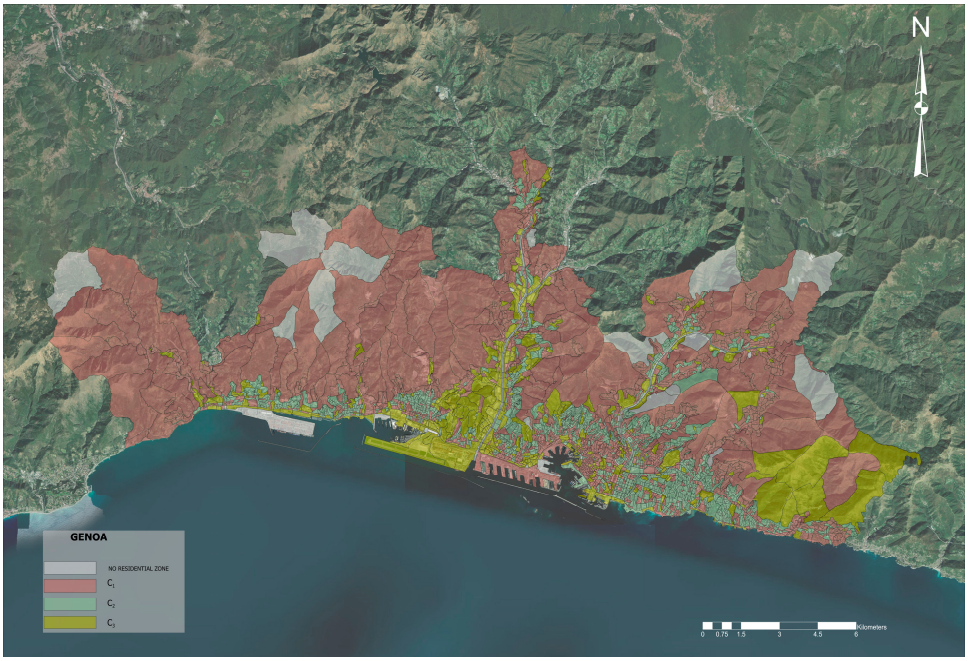

Figure 2. Thematic map of the census zones of Genoa classified into the tree clusters.

In summary, the urban zone of Genoa appears to be comprised of three distinct categories. The first category, depicted in red on the map, encompasses residential buildings predominantly constructed with load-bearing masonry and erected prior to the onset of the 20th century. These buildings are representative of the historic core of the city. The second category, represented by green on the map, is characterized by later urbanization, with buildings predominantly constructed using reinforced concrete. The third category, represented by orange on the map, comprises buildings that employ a combination of both construction techniques. These buildings were likely inhabited historically and have undergone subsequent urbanization.

4.2. Building Classification of Bari

Bari is partitioned in 1502 census zones of which 1450 are residential census zones (Figure 3).

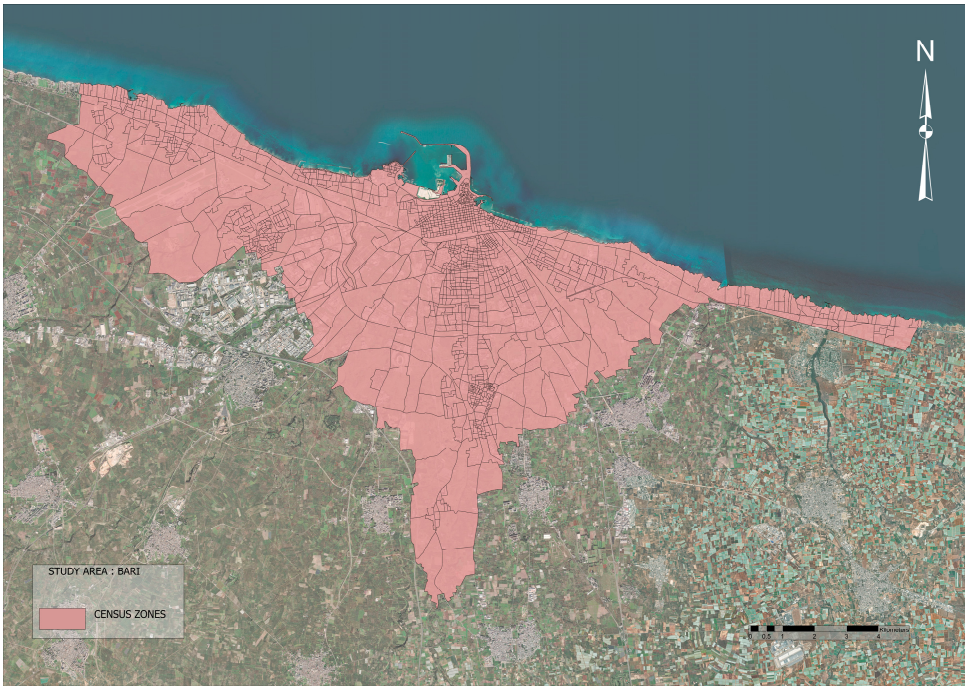

Figure 3. Partitioning of the city of Bari into census zones.

At the end of the preprocessing phase, the optimal number of clusters was obtained with value C = 2, for which the smallest value of the Xie–Beni index was measured to be equal to 0.225. For C = 2 the convergence is reached after six iterations, in which the value of the difference Δ in (17) is 3.426×10^{-6}.

Then, the census zones of Bari have been grouped into three clusters, C_1 and C_2. Table 5 shows, normalized with respect to the groups to which they belong, the centroids of the three clusters and the relevance of the features with respect to the three clusters.

Table 5. Cluster centroids and feature relevance for Bari.

Group	Feature	Cluster Centroids		Relevance	
		C_1	C_2	r_1	r_2
Construction technique	d_5	0.688	0.347	significant	not negligible
	d_6	0.259	0.603	negligible	not negligible
	d_7	0.053	0.051	negligible	negligible
Construction period	d_8	0.235	0.098	significant	negligible
	d_9	0.158	0.167	not negligible	not negligible
	d_{10}	0.161	0.173	not negligible	not negligible
	d_{11}	0.153	0.237	not negligible	significant
	d_{12}	0.180	0.205	not negligible	not negligible
	d_{13}	0.054	0.067	negligible	negligible
	d_{14}	0.037	0.036	negligible	negligible
	d_{15}	0.013	0.011	negligible	negligible
	d_{16}	0.009	0.005	negligible	negligible
Number of floors	d_{17}	0.202	0.100	negligible	negligible
	d_{18}	0.261	0.172	not negligible	negligible
	d_{19}	0.191	0.152	negligible	negligible
	d_{20}	0.347	0.576	not negligible	significant
Number of interiors	d_{21}	0.270	0.155	not negligible	negligible
	d_{22}	0.157	0.092	negligible	negligible
	d_{23}	0.151	0.107	negligible	negligible
	d_{24}	0.149	0.190	negligible	not negligible

https://doi.org/10.3390/sym18050807

	d_{25}	0.121	0.216	negligible	not negligible
	d_{26}	0.152	0.240	negligible	not negligible
State of conservation	d_{28}	0.248	0.174	negligible	negligible
	d_{29}	0.536	0.674	significant	significant
	d_{30}	0.195	0.143	negligible	negligible
	d_{31}	0.021	0.010	negligible	negligible

Cluster C1 comprises census zones that predominantly contain residential buildings constructed in load-bearing masonry prior to 1919 and that exhibit a satisfactory state of conservation. Cluster C2 comprises census zones that are characterized by the presence of residential buildings constructed in load-bearing masonry and reinforced concrete, with a frequency that is not negligible. The majority of these structures are in fair condition and were constructed primarily between 1919 and 1980, with a notable increase in the period between 1960 and 1970.

The thematic map in Figure 4 illustrates the partitioning of the census zones of Bari into three clusters. Non-residential census zones are indicated by the use of the color gray.

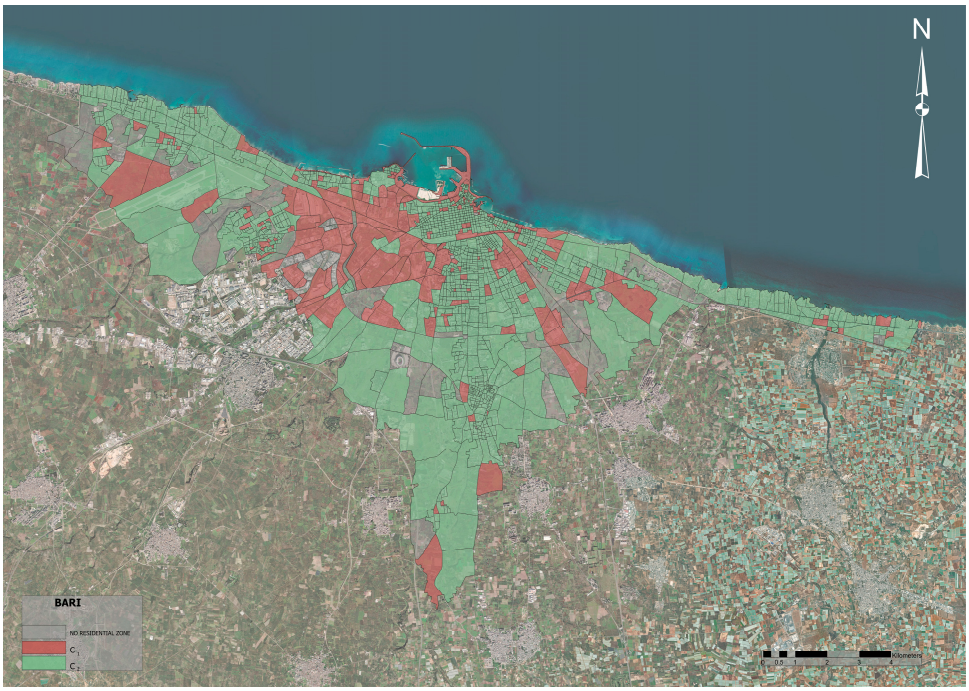

Figure 4. Thematic map of the census zones of Bari classified into the tree clusters.

Bari's urban landscape is characterized by the coexistence of two distinct residential zones, as depicted on the provided map. The first zone, delineated in red, comprises historic centers that have remained largely untouched by subsequent urban expansion. In contrast, the second zone, marked in green, consists of equally historic settlements that have undergone substantial building development between the post-war period and the 1980s. This phenomenon is exemplified by the coexistence within these urban areas of load-bearing masonry residential buildings, likely constructed between 1919 and 1945, and reinforced concrete buildings, presumably erected from the post-war period onwards.

4.3. Building Classification of Naples

Naples is partitioned in 4301 census zones of which 4049 are residential census zones (Figure 5).

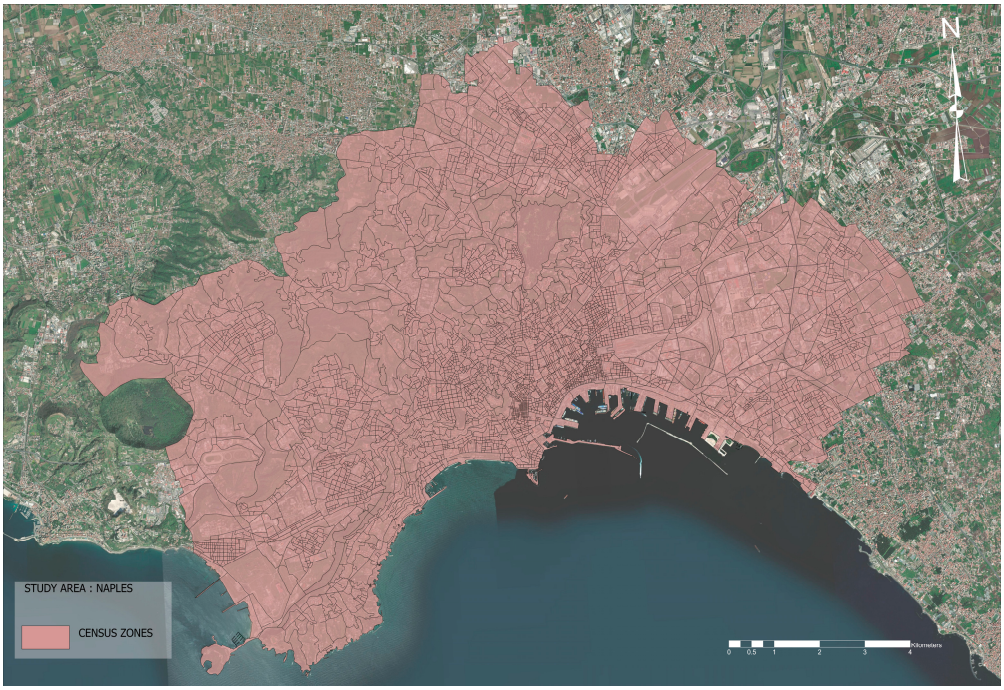

Figure 5. Partitioning of the city of Naples into census zones.

At the end of the preprocessing phase, the optimal number of clusters was obtained with value C = 3, for which the smallest value of the Xie–Beni index was measured to be equal to 2.480.

For C = 3 the convergence is reached after six iterations, in which the value of the difference Δ in (17) is 2534×10^{-6}.

Then, the census zones of Naples have been grouped into three clusters, C_1, C_2, and C_3. Table 6 shows, normalized with respect to the groups to which they belong, the centroids of the three clusters and the relevance of the features with respect to the three clusters.

Table 6. Cluster centroids and feature relevance for Naples.

Group	Feature	Cluster Centroids			Relevance		
		C_1	C_2	C_3	r_1	r_2	r_3
Construction technique	d_5	0.758	0.441	0.961	significant	not negligible	significant
	d_6	0.212	0.514	0.034	negligible	not negligible	negligible
	d_7	0.030	0.046	0.005	negligible	negligible	negligible
Construction period	d_8	0.415	0.200	0.769	significant	not negligible	significant
	d_9	0.212	0.139	0.173	not negligible	not negligible	not negligible
	d_{10}	0.134	0.197	0.021	not negligible	not negligible	negligible
	d_{11}	0.100	0.218	0.012	negligible	not negligible	negligible
	d_{12}	0.067	0.117	0.009	negligible	not negligible	negligible
	d_{13}	0.033	0.066	0.008	negligible	negligible	negligible
	d_{14}	0.006	0.015	0.000	negligible	negligible	negligible
	d_{15}	0.007	0.011	0.002	negligible	negligible	negligible
	d_{16}	0.025	0.037	0.006	negligible	negligible	negligible
Number of floors	d_{17}	0.075	0.087	0.053	negligible	negligible	negligible
	d_{18}	0.186	0.225	0.068	negligible	negligible	negligible
	d_{19}	0.142	0.138	0.060	negligible	negligible	negligible
	d_{20}	0.598	0.550	0.819	significant	significant	significant
Number of interiors	d_{21}	0.134	0.174	0.053	negligible	not negligible	negligible
	d_{22}	0.076	0.086	0.050	negligible	negligible	negligible

https://doi.org/10.3390/sym18050807

	d_{23}	0.126	0.114	0.148	negligible	negligible	negligible
	d_{24}	0.214	0.156	0.349	not negligible	negligible	significant
	d_{25}	0.220	0.143	0.282	not negligible	negligible	not negligible
	d_{26}	0.229	0.327	0.117	not negligible	not negligible	negligible
State of conservation	d_{28}	0.050	0.089	0.008	negligible	negligible	negligible
	d_{29}	0.473	0.562	0.163	not negligible	significant	negligible
	d_{30}	0.414	0.312	0.752	not negligible	not negligible	significant
	d_{31}	0.063	0.037	0.077	negligible	negligible	negligible

Cluster C_1 groups together census zones in which residential buildings were predominantly built in load-bearing masonry at a time before 1919, but with a non-negligible frequency of buildings constructed later, up until 1960. These residential buildings are, predominantly, at least four floors high. Cluster C_2 groups together census zones in which residential buildings were predominantly built in reinforced concrete in the post-war period, between 1945 and 1960, and in a good state of preservation. In this cluster too, residential buildings are predominantly at least four stories high.

Cluster C_3 includes census zones in which mainly residential buildings were constructed in load-bearing masonry before 1919. Most of them are in a poor state of preservation.

The thematic map in Figure 6 shows the partitioning of the census zones of Bari into the three clusters. Non-residential census zones are highlighted in grey.

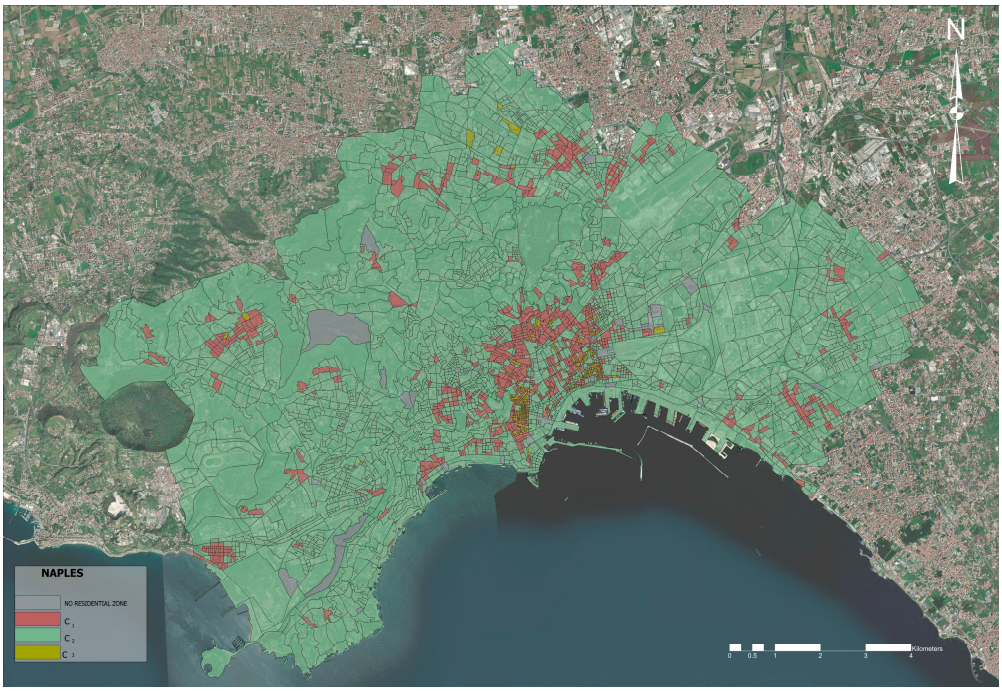

Figure 6. Thematic map of the census zones of Naples classified into the tree clusters.

In summary, Naples appears to be made up of three types of urban zones: those, displayed in red on the thematic map, with residential buildings predominantly made of load-bearing masonry and constructed before the beginning of the last century, which characterize the historic center of the city; they are, with high frequency, in poor state of conservation; those, displayed in green on the map, with later urbanization, with buildings constructed predominantly in reinforced concrete; and those, displayed in orange on the map, with buildings built predominantly in load-bearing masonry, but with subsequent constructions, probably in reinforced concrete, carried out up until 1960.

4.4. Comparison Results

The experimental comparison focuses on EWFCM, which is a representative state-of-the-art extension of FCM that incorporates entropy-regularized feature weighting. Comparative tests performed in [8] have shown that EWFCM consistently outperforms both the classical FCM and weighted FCM (wFCM) on standard benchmark datasets, including those from the UCI repository. Since EWFCM can be regarded as a generalization of these methods, comparing the proposed approach with EWFCM provides a stronger baseline.

These comparative tests have been carried by running WEFCM on all test cases. The stop iteration error ε in WEFCM is fixed to the value 1×10^{-5}.

Additionally, for EWFCM, 20 independent runs of each algorithm with different random initializations for each test case. The results are reported in terms of average.

First, the computational speeds of the two algorithms were measured. In Table 7 are shown the number of iterations and the CPU times obtained by running WEFCM and GWEFCM in the 16 test cases.

Table 7. Number of iterations and CPU times obtained executing WEFCM and GWEFCM.

City	Number of Clusters	WEFCM		GWEFCM	
		Iterations	CPU Time (s)	Iterations	CPU Time (s)
Tourin	3	15	17.41	7	3.28
Genoa	3	13	15.33	6	3.04
Milan	2	16	20.27	9	3.56
Venice	2	12	15.10	6	3.05
Verona	3	12	14.65	7	2.89
Padua	4	13	14.89	7	2.92
Parma	2	11	14.54	5	2.86
Bologna	3	14	16.76	7	3.10
Florence	2	14	16.32	6	3.05
Rome	2	16	20.06	8	3.55
Naples	3	15	18.60	7	3.37
Bari	2	11	15.24	6	3.07
Palermo	2	11	15.16	6	3.04
Catania	3	11	14.81	6	2.90
Messina	2	10	13.95	5	2.86
Cagliari	2	10	13.98	5	2.87

In all cases, GWEFCM reaches convergence in an average number of iterations equal to half the number of iterations of WEFCM, with CPU times on average equal to one third of those obtained by running WEFCM. These results highlight that GWEFCM is much more efficient than WEFCM in terms of execution times. This greater efficiency is since, unlike WEFCM, in which the feature weights vary within each cluster, in GWEFCM the feature weights do not vary across clusters; they are calculated as the product of the weight of the group to which they belong and the weight of the feature within the group.

To assess the statistical significance of the differences between GEWFCM and EWFCM, a Wilcoxon signed-rank test [22] was performed on the number of iterations and CPU execution times reported in Table 6. In the test the null hypothesis H_0 is that GEWFCM and EWFCM have the same performance. A significance level $\alpha = 0.05$ is set. The results of this test are shown in Table 8.

Table 8. Results of the Wilcoxon signed-rank test applied to number of iterations and CPU time.

Measure	*p*-Value
Iterations	1.53×10^{-5}
CPU time	1.53×10^{-5}

Therefore, the null hypothesis of equal performance between the two algorithms is rejected, indicating that the observed improvements are statistically significant and not due to random variability. This indicates that the observed improvements of GEWFCM in terms of number of iterations and CPU time are not due to random variability but reflect a systematic advantage of the proposed method.

Further comparative tests were performed to measure the similarity between the results obtained with the two algorithms. To measure this similarity, the Adjusted Rand Index (ARI) [23,24] was used; this index allows comparing the similarity between partitions obtained by two clustering algorithms. In Table 9 the values of the ARI metrics obtained in each test case are shown.

Table 9. Values of the ARI index for the 16 test cases.

City	ARI Index
Tourin	0.906
Genoa	0.871
Milan	0.862
Venice	0.915
Verona	0.877
Padua	0.904
Parma	0.917
Bologna	0.889
Florence	0.903
Rome	0.866
Naples	0.890
Bari	0.901
Palermo	0.905
Catania	0.903
Messina	0.912
Cagliari	0.910

The ARI values range between 0.861 and 0.917, with a mean of 0.896 and a standard deviation of 0.017. Given that ARI values lie in the interval [0, 1], these results demonstrate a strong agreement between the partitions produced by WEFCM and GWEFCM, confirming that the proposed method preserves the clustering structure of the baseline approach.

In summary, GWEFCM is comparable to WEFCM in terms of the quality of the clustering results, but it is computationally much faster.

Furthermore, it allows for assigning greater semantic meaning to clusters, highlighting how significant a feature is within a cluster compared to the group to which it belongs.

5. Conclusions

This study examined the most significant and representative feature-weighted fuzzy clustering algorithms. In most practical clustering tasks, attributes representing data characteristics have varying degrees of importance in forming the clustering structure.

This work presents a new entropy-weighted fuzzy clustering algorithm based on a two-level weight hierarchy: a global level that refers to groups of features and a global level that determines the weight of features within a group.

This mechanism, on the one hand, provides greater semantic interpretability of the clusters, and on the other, guarantees high computational speed, due to the independence of the weights from the clusters.

Experimental tests conducted on residential building census datasets from 16 Italian cities have shown that GEWFCM increases the semantic interpretability of the clusters. Furthermore, in all test cases, the results are highly similar to EWFCM, despite significantly lower processing times.

Further testing will be necessary to demonstrate the effectiveness of the proposed two-level hierarchy in optimally capturing the intrinsic structure of the data and providing better semantic interpretability of the clusters. To conclude, the authors intend to conduct future research to test GEWFCM on high-dimensional datasets with different cardinalities and data structures.

Author Contributions: Conceptualization, R.C., B.C., and F.D.M.; methodology, R.C., B.C., and F.D.M.; software, R.C., B.C., and F.D.M.; validation, R.C., B.C., and F.D.M.; formal analysis, R.C., B.C., and F.D.M.; investigation, R.C., B.C., and F.D.M.; resources, R.C., B.C., and F.D.M.; data curation, R.C., B.C., and F.D.M.; writing—original draft preparation, R.C., B.C., and F.D.M.; writing—review and editing, R.C., B.C., and F.D.M.; visualization, R.C., B.C., and F.D.M.; supervision, R.C., B.C., and F.D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study and the source code created to implement the proposed method are available on request from the corresponding author.

Acknowledgments: The article has been developed within the context of the project RETURN (Multi-Risk sciEnce for resilienT commUnities undeR a changiNg climate)—the extended partnership that aims to strengthen research chains on environmental, natural and anthropogenic risks at national level and promote their participation in strategic European and global value chains.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bezdek, J.C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Kluwer Academic Publishers: Norwell, MA, USA, 1981; p. 256. https://doi.org/10.1007/978-1-4757-0450-1.
2. Bezdek, J.C.; Ehrlich, R.; Full, W. FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci.* **1984**, *10*, 191–203. https://doi.org/10.1016/0098-3004(84)90020-7.
3. Li, W.; Zhai, S.; Xu, W.; Pedrycz, W.; Qian, Y.; Ding, W.; Zhan, T. Feature Selection Approach Based on Improved Fuzzy C-Means with Principle of Refined Justifiable Granularity. *IEEE Trans. Fuzzy Syst.* **2023**, *31*, 2112–2126. https://doi.org/10.1109/TFUZZ.2022.3217377.
4. Lincy, J.M.; Sudha, N. Weighted fuzzy C means and enhanced adaptive neuro-fuzzy inference based chronic kidney disease classification. *J. Fuzzy Ext. Appl.* **2024**, *5*, 100–115. https://doi.org/10.22105/jfea.2024.437690.1376.
5. Ma, N.; Hu, Q.; Wu, K.; Yuan, Y. A Dissimilarity Measure Powered Feature Weighted Fuzzy C-Means Algorithm for Gene Expression Data. *IEEE Trans. Fuzzy Syst.* **2025**, *33*, 192–202. https://doi.org/10.1109/TFUZZ.2024.3387465.
6. Keller, A.; Klawonn, F. Fuzzy clustering with weighting of data variables. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2000**, *8*, 734–746. https://doi.org/10.1142/S0218488500000538.
7. Winkler, R.; Klawonn, F.; Kruse, R. Fuzzy c-means in high dimensional spaces. *Int. J. Fuzzy Syst. Appl.* **2011**, *1*, 1–16. https://doi.org/10.4018/IJFSA.2011010101.
8. Zhou, J.; Chen, L.; Chen, C.I.P.; Zhang, Y.; Li, H.-Z. Fuzzy clustering with the entropy of attribute weights. *Neurocomputing* **2016**, *198*, 125–134. https://doi.org/10.1016/j.neucom.2015.09.127.

9. Yang, M.-S.; Nataliani, Y. A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy. *IEEE Trans. Fuzzy Syst.* **2017**, *26*, 817–835. https://doi.org/10.1109/TFUZZ.2017.2692203.
10. Guo, Y.; Wang, R.; Zhou, J.; Chen, Y.; Jiang, H.; Han, S.; Wang, L.; Du, T.; Ji, K.; Zhao, Y.; et al. Soft Subspace Fuzzy Clustering with Dimension Affinity Constraint. *Int. J. Fuzzy Syst.* **2022**, *24*, 2283–2301. https://doi.org/10.1007/s40815-022-01271-6.
11. Hashemzadeh, M.; Oskouei, A.G.; Farajzadeh, N. New fuzzy C-means clustering method based on feature-weight and cluster-weight learning. *Appl. Soft Comput.* **2019**, *78*, 324–345. https://doi.org/10.1016/j.asoc.2019.02.038.
12. Song, S.; Jia, Z.; Shi, F.; Wang, J.; Ni, D. Adaptive fuzzy weighted C-mean image segmentation algorithm combining a new distance metric and prior entropy. *Eng. Appl. Artif. Intell.* **2024**, *131*, 107776. https://doi.org/10.1016/j.engappai.2023.107776.
13. Lin, J.; Wu, L.; Chen, R.; Wu, J.; Wang, X. Double-weighted fuzzy clustering with samples and generalized entropy features. *Concurr. Comput. Pract. Exp.* **2021**, *33*, e5758. https://doi.org/10.1002/cpe.5758.
14. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006; 778p, ISBN 978-0-387-31073-2.
15. He, Y.-L.; Zhang, X.-L.; Ao, W.; Huang, J.Z. Determining the optimal temperature parameter for softmax function in reinforcement learning. *Appl. Soft Comput.* **2018**, *70*, 80–85. https://doi.org/10.1016/j.asoc.2018.05.012.
16. Amini Pishro, A.; Zhang, S.; L◙Hostis, A.; Liu, Y.; Hu, Q.; Hejazi, F.; Shahpasand, M.; Rahman, A.; Oueslati, A.; Zhang, Z. Machine learning-aided hybrid technique for dynamics of rail transit stations classification: A case study. *Sci. Rep.* **2024**, *14*, 23929. https://doi.org/10.1038/s41598-024-75541-8.
17. Cafaro, R.; Cardone, B.; D◙Ambrosio, V.; Di Martino, F.; Miraglia, V. A GIS-Integrated Framework for Unsupervised Fuzzy Classification of Residential Building Pattern. *Electronics* **2025**, *14*, 4022. https://doi.org/10.3390/electronics14204022.
18. Xie, X.L.; Beni, I.G. A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*, 841–847 https://doi.org/10.1109/34.85677.
19. Pal, N.R.; Bezdeck, J.C. On cluster validity for the fuzzy c-means model. *IEEE Trans. Fuzzy Syst.* **1995**, *3*, 370–379. https://doi.org/10.1109/91.413225.
20. Pérez-Sánchez, I.; Medina-Pérez, M.A.; Monroy, R.; Loyola-González, O.; Gutierrez-Rodríguez, A.E. New Evaluation Method for Fuzzy Cluster Validity Indices. *IEEE Access* **2025**, *13*, 22728–22744 https://doi.org/10.1109/ACCESS.2025.3535417.
21. Huang, M.; Xia, Z.; Wang, H.; Zeng, Q.; Wang, Q. The range of the value for the fuzzifier of the fuzzy c-means algorithm. *Pattern Recognit. Lett.* **2012**, *33*, 2280–2284. https://doi.org/10.1016/j.patrec.2012.08.014.
22. Conover, W.J. *Practical Nonparametric Statistics*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 1999; 608p, ISBN 978-0-471-16068-7.
23. Santos, J.M.; Embrechts, M. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. In *Artificial Neural Networks–ICANN 2009. ICANN 2009*; Alippi, C., Polycarpou, M., Panayiotou, C., Ellinas, G., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5769. https://doi.org/10.1007/978-3-642-04277-5_18.
24. Warrens, M.J.; van der Hoef, H. Understanding the Adjusted Rand Index and Other Partition Comparison Indices Based on Counting Object Pairs. *J. Classif.* **2022**, *39*, 487–509. https://doi.org/10.1007/s00357-022-09413-z.