Graph-Based Clustering of Urban Water Consumption Profiles via Adaptive Attention and Multi-Relational Topologies

Jonnatan Arias García , David Cárdenas-Peña * , Alvaro Orozco-Gutierrez , Hernan Felipe Garcia-Arias , Jhoniers Gilberto Guerrero-Erazo

Posted Date: 1 April 2026

doi: 10.20944/preprints202604.0031.v1

Keywords: graph autoencoders; graph neural networks (GNNs); socio-demographic segmentation; spatio-temporal clustering; urban resource management; water consumption analysis

 doi:10.20944/preprints202604.0031.v1

Article

Graph-Based Clustering of Urban Water Consumption Profiles via Adaptive Attention and Multi-Relational Topologies

Jonnatan Arias-Garcia [1], David Cárdenas-Peña [1,*], Alvaro Angel Orozco-Gutiérrez [1], Hernan Felipe Garcia-Arias [2] and Jhoniers Gilberto Guerrero-Erazo [3]

[1] Automatics Research Group, Universidad Tecnológica de Pereira, Pereira 660003, Colombia
[2] SISTEMIC Research Group, Universidad de Antioquia, Medellín 050010, Colombia
[3] Water and Sanitation Research Group, Universidad Tecnológica de Pereira, Pereira, Colombia
* Correspondence: dcardenasp@utp.edu.co

Abstract

Conventional clustering techniques for urban water consumption profiling treat each household as an independent entity, thereby disregarding the spatial, socioeconomic, and infrastructural contexts that jointly govern demand behavior. This structural limitation prevents the extraction of contextually coherent consumption profiles—a critical shortcoming for utility managers who must design spatially targeted conservation interventions. To overcome this, we propose Simple GLAC, a novel graph clustering framework that leverages graph neural networks with an adaptive attention mechanism to dynamically model these complex interdependencies. The model's end-to-end training jointly optimizes a latent representation for cluster cohesion, separation, and spatial homogeneity, where each household's multi-month consumption record serves as the node feature vector encoding temporal consumption patterns. Evaluated on a large-scale real-world dataset of 4590 residential households across four distinct graph topologies, Simple GLAC consistently achieves superior multi-metric performance over both traditional and graph-based benchmarks, yielding interpretable and operationally actionable consumption profiles aligned with the spatial, administrative, socioeconomic, and infrastructural dimensions of urban water governance. This work provides a powerful, data-driven tool for utility managers to deploy targeted water conservation strategies and optimize urban resource distribution.

Keywords: graph autoencoders; graph neural networks (GNNs); socio-demographic segmentation; spatio-temporal clustering; urban resource management; water consumption analysis

1. Introduction

Urban water systems face increasing pressure from urbanization and climate variability, demanding a shift from aggregate demand forecasting toward behavior-based consumption profiling [1]. Then, identifying distinct household usage patterns can inform targeted conservation strategies, equitable tariff design, and infrastructure prioritization. Clustering methodologies address this need by grouping households based on the shape and magnitude of their consumption time series and have demonstrated practical utility in smart meter deployments worldwide [2]. However, traditional clustering approaches treat each household as an isolated statistical entity, analyzing consumption trajectories without regard for the relational context surrounding it, such as shared socioeconomic attributes or common hydraulic infrastructure [3]. Although capturing the magnitude and periodicity of consumption, such approaches remain structurally blind underlying factors that determine where and why those patterns emerge, resulting in a limitation that prevents the extraction of profiles that are both statistically compact and contextually interpretable [4].

Graph-based learning provides a principled framework for overcoming the above limitation by representing households as nodes and their complex relationships as weighted edges, explicitly

 doi:10.20944/preprints202604.0031.v1

encoding the multi-dimensional relational context that governs consumption behavior [5]. Graph neural networks (GNNs) then propagate this contextual information via message passing, enabling each household's representation to be enriched by its neighbors' attributes and structural positions [6]. Conventionally, the message passing is driven by reconstruction fidelity, node classification accuracy, or the goal of producing pure, spatially coherent, and semantically interpretable consumption clusters [7]. However, a fundamental tension remains between the above training objectives. Moreover, most existing graph clustering approaches either decouple the representation learning step from the clustering objective, leading to suboptimal latent spaces, or rely on fixed, predefined graph structures that cannot adaptively refine edge weights based on the latent consumption similarities that emerge during training [8].

Graph-based learning provides a principled framework for overcoming the above limitation by representing households as nodes and their complex relationships as weighted edges, explicitly encoding the multi-dimensional relational context that governs consumption behavior [5]. Graph neural networks (GNNs) then propagate this contextual information via message passing, enabling each household's representation to be enriched by its neighbors' attributes and structural positions [6]. Conventionally, the message passing is driven by reconstruction fidelity, node classification accuracy, or the goal of producing pure, spatially coherent, and semantically interpretable consumption clusters. However, a fundamental tension remains between the above training objectives. Moreover, most existing graph clustering approaches either decouple the representation learning step from the clustering objective, leading to suboptimal latent spaces, or rely on fixed, predefined graph structures that cannot adaptively refine edge weights based on the latent consumption similarities that emerge during training [8]. In this context, incorporating supervised or similarity-aware mechanisms during representation learning has been shown to enhance the discriminative structure of latent spaces and improve performance [9].

To address these limitations, this work proposes an end-to-end graph learning with an adaptive clustering, termed Simple GLAC, that integrates a trainable edge attention mechanism into a graph convolutional encoder and minimizes a composite loss function that jointly enforces spatial homogeneity, cluster assignment confidence, size balance, and centroid separation. Each household's multi-month consumption record constitutes the node feature vector, while the graph topology encodes relational context through four distinct formulations. The model is rigorously evaluated across four graph topologies (geospatial, territorial, socioeconomic, and hydraulic) constructed from a large-scale real-world dataset of 4,590 residential households in Pereira, Colombia. The principal contributions of this work are threefold: the introduction of an adaptive attention mechanism that dynamically refines edge weights during end-to-end training; a composite loss function that simultaneously optimizes spatial homogeneity, assignment confidence, cluster balance, and centroid separation; and a systematic comparative analysis against baseline methods across all four topologies, demonstrating that the relational topology is a determinant of cluster semantics and that Simple GLAC constitutes the most robust method under a multi-metric evaluation framework.

The remainder of this work is organized as follows. Section 2 reviews the related literature and establishes the context for graph-based clustering approaches. Section 3 describes the dataset, graph construction strategies, and the proposed Simple GLAC methodology, along with the experimental setup. Section 4 presents and discusses the quantitative and qualitative results, including cluster-level analysis across different graph topologies. Finally, Section 5 concludes the paper and outlines limitations and directions for future research.

2. Related Work

The analysis of public utilities consumption, particularly water demand, has historically been dominated by time series clustering techniques applied to smart meter data, with established algorithms such as K-means [10], Hierarchical clustering [11], and Gaussian Mixture Models (GMMs) [12] serving as the foundational toolkit for segmenting consumers into distinct behavioral profiles

 doi:10.20944/preprints202604.0031.v1

[13,14]. These methodologies excel at identifying temporal regularities, such as diurnal and seasonal cycles, by operating on consumption vectors in isolation, effectively grouping households based on the magnitude and periodicity of their water use [15]. However, a significant limitation of these conventional approaches is their treatment of each household as an independent entity, analyzing consumption trajectories without regard for the rich contextual information surrounding it. Hence, these methods successfully characterize the temporal regularity of individual consumption (e.g., magnitude, periodicity, and seasonal variation) while remaining structurally agnostic to the spatial dependencies, socioeconomic influences, and infrastructure network that collectively determine where and why those patterns emerge [16,17]. For instance, such approaches assign two households with identical consumption time series to the same group even if they belong to fundamentally different consumption profiles: one is located in a high-income district with water-intensive practices, and the other in a low-income area with a shared connection subject to pressure intermittency.

This methodological gap has catalyzed a paradigm shift towards graph-based learning frameworks, which offer a natural, mathematically rigorous substrate for integrating heterogeneous data types [18,19]. Specifically, by representing households as nodes and their complex interrelationships (e.g., geographical proximity or administrative boundaries) as edges, graph structures inherently encode the multifaceted context of consumption [20–22]. Within such a graph-structured domain, the problem of profiling water consumption can be conceptually decomposed into two interconnected challenges. The first is information diffusion, which concerns how node features (consumption data) propagate and refine through the graph topology to yield context-aware representations; the second is node clustering, which entails partitioning these refined representations into coherent and meaningful consumption profiles [23,24].

Seminal approaches to graph-based consumption analysis primarily leveraged classic Graph Convolutional Networks (GCNs) [25,26] for the diffusion (message-passing) step, followed by the application of standard clustering algorithms, such as GMMs, to the resulting node embeddings [27,28]. While this two-stage pipeline was effective, it often suffered from a disconnect between the objective of the graph network—typically trained for reconstruction or a surrogate task—and the ultimate goal of cluster purity. To address this disconnect, autoencoder-based architectures were introduced for learning a unified latent space that jointly preserves graph structure and node attributes. Building on this development, models such as the Graph Autoencoder with Gaussian Mixture Model (AE-G-GMM) [29–31] attempt to combine the representational power of graph convolutional encoders with the probabilistic assignment of GMMs, directly optimizing the latent embeddings for clustering fidelity. Nevertheless, these methods are constrained by their reliance on a fixed, predefined graph structure, which cannot adapt to latent consumption similarities discovered during training. Moreover, the autoencoder's reconstruction objective is not aligned with cluster purity: two structurally similar neighborhoods (i.e., nodes with comparable adjacency patterns) may exhibit divergent consumption behaviors due to socioeconomic or infrastructural factors that the fixed topology fails to capture. This mismatch between structural similarity and behavioral similarity is especially pronounced in heterogeneous urban water networks, where geographically proximate households may belong to different socioeconomic strata or hydraulic sectors.

The pursuit of more adaptive, topologically aware models has led to novel architectures for graph clustering. Methods such as GLAC-GCN and Graph Multiview introduce significant advancements by using global and local topology-aware contrastive learning and enabling the model to capture nuanced cluster structures beyond neighborhood smoothing [32,33]. Their specialized loss functions promote intra-cluster cohesion and inter-cluster separation. Despite these advances, three gaps remain unaddressed. First, no prior work has evaluated graph clustering for water consumption profiling across multiple heterogeneous graph topologies simultaneously. Most studies use a single relational encoding and cannot assess how topology choice affects cluster semantics. Second, existing composite loss functions do not explicitly penalize cluster-size imbalance. This is critical in urban water datasets, where a small number of high-consumption outlier households can dominate centroid positioning.

 doi:10.20944/preprints202604.0031.v1

Third, adaptive refinement of edge weights during end-to-end clustering training, instead of fixing them before training, has not been systematically explored for consumption profiling. The proposed Simple GLAC model is designed to address all three gaps.

3. Materials and Methods

3.1. Dataset and Data Pre-Processing

This study utilizes a comprehensive dataset of water consumption information from residential households in Pereira, Colombia, comprising a statistically representative subset of 4590 households selected through stratified sampling techniques that preserve diversity while maintaining computational feasibility. The dataset encompasses multiple dimensions of urban water usage:

- Geospatial features, comprising latitude, longitude, neighborhood, and administrative zone, enable the granular analysis of water consumption patterns across different regions within the city.
- Socioeconomic attributes support the analysis of the relationship between economic status and water usage in terms of the stratum (a classification system in Colombia that categorizes households based on income and living conditions) and corresponding water tariffs ($\$COP/m^3$).
- Hydraulic sector identification defines the node within the water distribution network to which a household is assigned, providing information about the infrastructure and water distribution practices.
- Monthly water consumption data with records spanning six years, from 2017 to early 2023, yielding a timeseries with 78 time instants (months). The average water consumption values within the analysis period is $11.7 m^3$/month/household, with a variance of 65.6. Along with consumptions values in the interval $[0.0, 271.0] m^3$, above statistics indicate moderate dispersion and the presence of extreme values, likely influencing conventional clustering approaches.

To prepare the dataset, data preprocessing involved a systematic process to ensure data quality and reliability. The process began with the temporal alignment of consumption records for all households. Initial statistical analysis identified a few households with abnormally high consumption patterns, which corresponded to commercial users misclassified as residential, and these were excluded from the study. For missing values in each household's monthly consumption record, we evaluated two imputation strategies: (i) replacement by the household's temporal mean, and (ii) k-nearest neighbor (KNN) imputation. KNN imputation was performed in the temporal feature space—using the full 78-month consumption vector as the similarity basis with Euclidean distance—identifying the five most similar households (k = 5) for each record with missing entries and imputing the missing month with the average of the corresponding month across those five neighbors. Only households with fewer than 10% missing values (i.e., fewer than 8 months) were retained prior to imputation; those exceeding this threshold were excluded. The 5-NN strategy was selected over mean imputation based on lower reconstruction error on a held-out validation subset of complete records with simulated missingness at random.

3.2. Simple Graph Learning with Adaptive Clustering (Simple GLAC)

To address the complex task of identifying water consumption profiles by integrating geospatial, socioeconomic, and hydraulic data, we propose a novel graph-based clustering model termed Simple GLAC (Graph Learning with Adaptive Clustering). This model is designed to learn cohesive cluster assignments directly within a latent representation space, obviating the need for a separate, post-hoc clustering step. The methodology unfolds in two primary phases: the construction of a heterogeneous graph from multi-source household data, followed by the application of the end-to-end Simple GLAC architecture.

The foundational element of the proposed approach is the weighted graph denoted in Equation (1)

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{W}) \tag{1}$$

 doi:10.20944/preprints202604.0031.v1

where $\mathcal{E}$ denotes the edges structuring the graph through adjacencies among nodes in the set $\mathcal{V}$:

$$\mathcal{E} = \{(i,j) \mid \text{dist}(v_i, v_j) \leq \epsilon \; \forall i,j \in \mathcal{V}\}, \tag{2}$$

with $\text{dist}(\cdot,\cdot)$ as a distance function and $\epsilon \in \mathbb{R}^+$ a threshold truncating adjacency. The node feature matrix $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times d}$ encapsulates the feature vector $\mathbf{x}_i \in \mathbb{R}^d$ representing the local attributes of the node $i \in \mathcal{V}$. The relational structure is encoded in a weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, with elements w_{ij} quantifying the connection strength between nodes i and j. These edge weights are computed as a convex combination of spatial and temporal similarities formalized by the Equation (3).

$$w_{ij} = \begin{cases} \alpha \cdot \text{Spatial Similarity}(i,j) + & \text{if } (i,j) \in \mathcal{E} \\ (1-\alpha) \cdot \text{Consumption Similarity}(i,j), & \\ 0, & \text{if } (i,j) \notin \mathcal{E} \end{cases} \tag{3}$$

where $\alpha \in (0,1)$ is a hyperparameter controlling the trade-off between geographic proximity and consumption pattern similarity, respectively. The euclidean distance between geographic coordinates (latitude and longitude) $\boldsymbol{s}_i, \boldsymbol{s}_j \in \mathbb{R}^2$ assesses the $\text{Spatial Similarity}(i,j)$ as:

$$\text{Spatial Similarity}(i,j) = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|, \tag{4}$$

In turn, the cosine similarity between the respective consumption time series $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^d$ measures the $\text{Consumption Similarity}(i,j)$, capturing behavioral synchronicity independent of absolute magnitude:

$$\text{Consumption Similarity}(i,j) = \frac{\boldsymbol{x}_i{}^\top \boldsymbol{x}_i}{\|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|}. \tag{5}$$

The Simple Graph Learning with Adaptive Clustering (Simple GLAC) model works as an encoder network that transforms input node features into a latent space conducive to clustering. This encoder employs a dual-layer Graph Convolutional Network (GCN) architecture, enhanced with a dedicated edge attention mechanism. The forward propagation begins with a first graph convolutional block:

$$\mathbf{H} = \text{LeakyReLU}(\text{BatchNorm}(\text{GCNConv}(\mathcal{E}, \mathbf{X}, \mathbf{W} | \boldsymbol{\phi}_h))), \tag{6}$$

where the GCNConv operation performs a first-order neighborhood aggregation, the $\text{BatchNorm}(\cdot)$ layer stabilizes training by normalizing activations, and the LeakyReLU activation function introduces non-linearity. The resulting matrix $\mathbf{H} \in \mathbb{R}^{|\mathcal{V}| \times d_h}$ corresponds to the updated d_h features for each node. Subsequently, Simple GLAC adapts the edge weights in the original graph structure as follows:

$$w_{ij}^{(h)} = \begin{cases} \sigma\left(\boldsymbol{h}_i^\top \boldsymbol{\Theta} \boldsymbol{h}_j\right), & \text{if } (i,j) \in \mathcal{E} \\ 0, & \text{if } (i,j) \notin \mathcal{E} \end{cases} \tag{7}$$

with vectors $\boldsymbol{h}_i, \boldsymbol{h}_j \in \mathbb{R}^{d_h}$ denoting the d_h hidden features of nodes i and j, the diagonal matrix $\boldsymbol{\Theta} = \text{diag}\left(\{\theta_d\}_{d=1}^{d_h}\right)$ acting as a trainable feature-wise filter, and the sigmoid function $\sigma(\cdot)$ bounding the attention coefficient $w_{ij}^{(h)}$ within $[0,1]$ to indicate the strength of the adaptive relationship between nodes. Note that Simple GLAC refines the edge weights in Equation (3) through element-wise multiplication with $w_{ij}^{(h)}$. Resulting node features $\mathbf{H}$ and attention weights $\mathbf{W}^{(h)}$ feed a second graph convolutional block:

$$\mathbf{Z} = \text{BatchNorm}\left(\text{GCNConv}\left(\mathcal{E}, \mathbf{H}, \mathbf{W}^{(h)} | \boldsymbol{\phi}_z\right)\right) \tag{8}$$

The output of this block, $\mathbf{Z} \in \mathbb{R}^{|\mathcal{V}| \times d_z}$, serves as the final node embedding, which simultaneously captures the intrinsic attributes of each household and its contextual relationships within the attentively-weighted graph.

 doi:10.20944/preprints202604.0031.v1

Directly upon this $\mathbf{Z}$ embedding, cluster assignment is performed using a probabilistic approach based on the Student's t-distribution [34]. This methodological choice is motivated by both theoretical considerations and empirical evidence from deep clustering literature. The assignment probability of node i to cluster k is given by:

$$q_{ik} = \frac{\left(1 + \frac{||\mathbf{z}_i - \boldsymbol{\mu}_k||^2}{\nu}\right)^{\frac{\nu+1}{2}}}{\sum_{k'=1}^{K}\left(1 + \frac{||\mathbf{z}_i - \boldsymbol{\mu}_{k'}||^2}{\nu}\right)^{\frac{\nu+1}{2}}}, \tag{9}$$

being $\mathbf{z}_i \in \mathbb{R}^{d_z}$ the i-th node embedding, $\nu \in \mathbb{R}^+$ the degree of freedom of the t-distribution (usually 1, to manage outliers), $\boldsymbol{\mu}_k$ the trainable centroid vector for cluster k, and K the predefined number of clusters [35]. This probabilistic formulation leverages the t-distribution's heavy-tailed properties to mitigate outlier influence on centroid positioning, which is critical for noisy urban datasets. Simultaneously, Equation (9) establishes a differentiable clustering objective that enables joint refinement of embeddings and centroids through gradient-based optimization, following established practices in deep embedded clustering like GLAC-GCN [32].

An end-to-end model training is carried out by minimizing the multi-objective loss function in Equation (10) w.r.t. the model parameters $\{\boldsymbol{\phi}_h, \boldsymbol{\Theta}, \boldsymbol{\phi}_z\}$,

$$\mathcal{L}(\boldsymbol{\phi}_h, \boldsymbol{\Theta}, \boldsymbol{\phi}_z) = \lambda\, \mathcal{L}_{\text{spatial}} + (1-\lambda)\, \mathcal{L}_{\text{KL}} + \eta\, \mathcal{L}_{\text{balance}} + \xi\, \mathcal{L}_{\text{sep}} \tag{10}$$

$$\mathcal{L}_{\text{spatial}} = \mathbb{E}_{(i,j)\in\mathcal{E}}\left\{ w_{ij}^{(h)} \left\|\mathbf{z}_i - \mathbf{z}_j\right\|^2 \right\} \tag{11}$$

$$\mathcal{L}_{\text{KL}} = \text{KL}(q||p) \tag{12}$$

$$p(k|\mathbf{z}_i) = \frac{q(k|\mathbf{z}_i)^2 / \sum_i q(k|\mathbf{z}_i)}{\sum_{k'} q(k'|\mathbf{z}_i)^2 / \sum_i q(k'|\mathbf{z}_i)} \tag{13}$$

$$\mathcal{L}_{\text{balance}} = \frac{\text{std}(|\mathcal{C}_1| \ldots, |\mathcal{C}_K|)}{\text{mean}(|\mathcal{C}_1|, \ldots, |\mathcal{C}_K|)} \tag{14}$$

$$\mathcal{L}_{\text{sep}} = \frac{-2}{K(K-1)} \sum_{k<k'} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|^2 \tag{15}$$

with $\mathbb{E}\{\cdot\}$ as the expectation operator. The spatial loss term $\mathcal{L}_{\text{spatial}}$ encourages neighboring nodes in the original graph, particularly those with strong connections, to achieve similar embeddings in the latent space, thereby promoting spatial affinity. The clustering refinement loss $\mathcal{L}_{\text{KL}}$ corresponds to the Kullback-Leibler divergence between the assignment distribution $q(k|\mathbf{z}_i)$ computed according to Equation (9) and the target distribution $p(k|\mathbf{z}_i)$ in Equation (13) which enforces the assignment confidence. The cluster balance term $\mathcal{L}_{\text{balance}}$ uses the cluster cardinalities $|\mathcal{C}_k| \forall k \in \{1, \ldots, K\}$ to penalize significantly different group sizes, preventing degenerate assignments where most nodes are grouped into a single cluster. Lastly, the cluster separation loss $\mathcal{L}_{\text{sep}}$ directly acts on the trainable centroids to maximize the pairwise distance between them, ensuring that the learned consumption profiles are distinct and well-separated in the latent space. The hyperparameters $\lambda, \eta, \xi \in \mathbb{R}^+$ in Equation (10) trade-off the above competing objectives. Therefore, the synergistic optimization of this composite loss function makes the Simple GLAC model to generate a latent representation where fundamental water consumption profiles emerge naturally, cohesively, and inherently interpretable.

3.3. Experimental Framework

The experimental design of this work relies on a comparison against baseline approaches in terms of well-known clustering metrics, while varying graph structure. In the hand of traditional baseline approaches include: K-means with K-means++ initialization, seeking compact and spherical clusters; Hierarchical clustering (HC) with Ward's minimum variance method, minimizing intra-cluster variance; and Gaussian Mixture Models (GMM) with full covariance, allowing flexible cluster shapes. In the hand of graph-based approaches, four well-known approaches are contrasted: The straightforward

 doi:10.20944/preprints202604.0031.v1

graph convolutional network with GMM (GConv-GMM) [25] and the efficient graph convolutional network for clustering (ClusterGCN) [26] perform neighborhood aggregation using fixed graph structures. The autoencoder-based geometric deep learning (AE-G-GMM) combines graph convolutional encoders with Gaussian mixture models for joint representation learning and clustering [29]. The global and local topology-aware contrastive GCN (GLAC-GCN) incorporates attention mechanisms to dynamically weight neighborhood influences during information propagation [32].

Regarding the metrics, classical cluster validation metrics and benchmark measures quantify the quality of the obtained clusters. Traditional metrics, providing a foundational assessment, include: the silhouette score (SS), bounded by $[-1, 1]$, which measures point-cluster fit relative to neighboring clusters; the Davies-Bouldin index (DBS), to account for intracluster dispersion and low centroid separation [36]; and the Calinski-Harabasz index (CHI), computed as the ratio of intercluster to intracluster variance [37]. Two benchmark metrics are considered to mitigate the large-cluster bias and convexity assumption of CHI and DB. Firstly, the intra-cluster coherence (IC) is computed cluster-wise as in Equation (16), where $\mathcal{C}_k$ stands for the subset of points assigned to the k-th cluster. IC measures the distance between points and their assigned centroid $\boldsymbol{\mu}_k$, so that the lower the IC, the tighter and the more homogeneous the clusters. Secondly, the inter-cluster dissimilarity (ID), in Equation (17), corresponds to the average pairwise distance between centroids achieving large scores at better-separated groups. The joint analysis of Silhouette, Calinski–Harabasz, Davies–Bouldin scores, and Intercluster Dissimilarity allows to thoroughly evaluate the cluster separation, compactness, and centroid geometry.

$$\mathrm{IC}_k = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \|\boldsymbol{z}_i - \boldsymbol{\mu}_k\|^2 \tag{16}$$

$$\mathrm{ID} = \frac{2}{k(k-1)} \sum_{i<j} \|\mu_i - \mu_j\| \tag{17}$$

For graph construction, both the node feature matrix $\mathbf{X}$ and the edge set $\mathcal{E}$ must be defined. The feature matrix is constructed from the household water consumption time series, resulting in $|\mathcal{V}| = 4590$ nodes and $d = 78$ temporal features per node. Since the graph structure $\mathcal{E}$ governs the message-passing (diffusion) process, the experimental framework considers four distinct relational topologies, each encoding a different notion of household connectivity, as illustrated in Figure 1.

$\mathcal{E}_G$: The Geospatial graph, in Figure 1a, connects households based on spatial proximity using a k-nearest neighbors (KNN) scheme, where $\mathbf{v}_i = \mathbf{s}_i \in \mathbb{R}^2$ denotes geographic coordinates. The number of neighbors is set to $k = 50$ to ensure a sufficiently dense graph, enabling the capture of local spatial correlations independently of administrative or infrastructural constraints. Since this graph lacks an explicit grouping variable, Figure 1 illustrates only pairwise geographic relationships rather than aggregated consumption patterns.

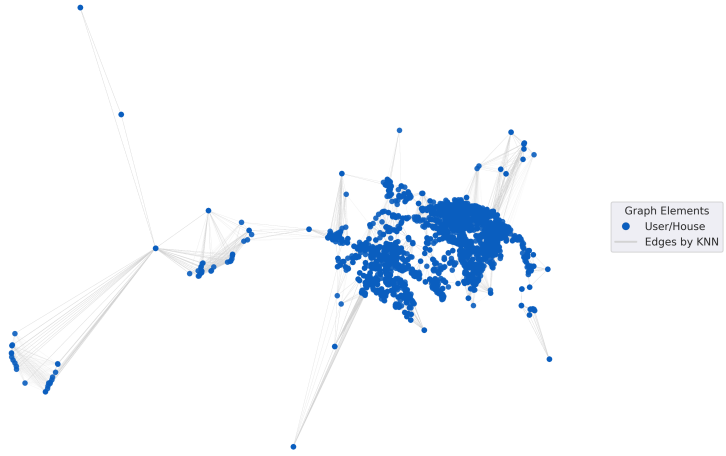
$\mathcal{E}_T$: In the Territorial graph, illustrated in Figure 1b, households are grouped according to administrative divisions (29 communes), such that $\mathbf{v}_i = \text{commune}_i$ and edges are restricted within each region ($\epsilon = 0$). This structure introduces explicit diffusion barriers aligned with urban governance boundaries (see Figure 1b). Further, the initial statistics (scatterplot in Figure 1b) reveal substantial spatial heterogeneity in consumption across zones, with average values ranging from approximately $6m^3$ to $20m^3$. High-density communes (e.g., C2, C24, C11) concentrate around $10\tilde{}11m^3$, forming a stable baseline of urban demand, while smaller or peripheral zones exhibit greater variability. These patterns highlight that consumption is strongly conditioned by administrative context, which is not captured by traditional clustering methods operating on independent time series. Therefore, the graph formulation must enable the integration of governance-aware structure into the clustering process, transforming purely statistical groupings into administratively meaningful profiles.

$\mathcal{E}_S$: The Socioeconomic graph connects households within the same socioeconomic stratum, defined by the Colombian six-level classification system ($\mathbf{v}_i \in 1, \ldots, 6, \epsilon = 0$), under the assumption that

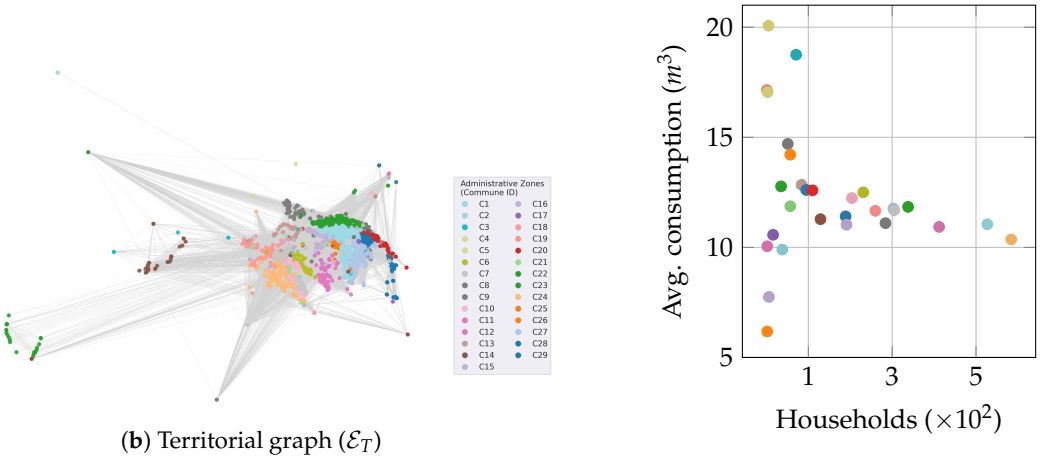
 doi:10.20944/preprints202604.0031.v1

income-related factors influence water consumption behavior (see Figure 1c). The consumption distribution (right panel) indicates a non-monotonic relationship between socioeconomic level and consumption. While Stratum 6 exhibits the highest mean consumption ($\approx 13.4m^3$) and large variance, indicating heterogeneity associated with high-income households, Stratum 1 also shows relatively elevated consumption ($\approx 12.1m^3$), likely driven by higher population density. In contrast, intermediate strata (4–5) display lower average consumption ($\approx 10.4 - 10.8m^3$), suggesting more efficient usage patterns. This irregular structure represents a key challenge for clustering, as consumption does not follow a simple socioeconomic gradient. Consequently, the effectiveness of each method depends on its ability to translate these heterogeneous patterns into coherent clusters through graph diffusion.

$\mathcal{E}_H$: The Hydraulic sector graph encodes connectivity based on the water distribution network, shown in Figure 1d, linking households according to their associated pipeline sector (see Figure 1d). This results in a structurally homogeneous topology in which nodes are grouped by shared infrastructure. Consumption statistics (scatterplot on the right) indicate relatively small differences in average consumption across sectors, with most values clustered around the citywide mean ($\approx 11.6m^3$). For instance, sector HS4 shows the highest consumption ($\approx 13.45m^3$) and variability, suggesting heterogeneous land use or infrastructure conditions, while HS5 exhibits the lowest consumption and variance, indicating more uniform demand. The remaining sectors (HS1–HS3) lie within a narrow range ($\approx 11.3 - 11.8m^3$), reflecting limited inter-group contrast. This infrastructure-based topology introduces a diffusion regime dominated by homogenization, where strong connectivity tends to smooth consumption differences. As a result, the main challenge is not identifying large-scale variations, but preserving meaningful deviations from the mean while maintaining coherent clusters.



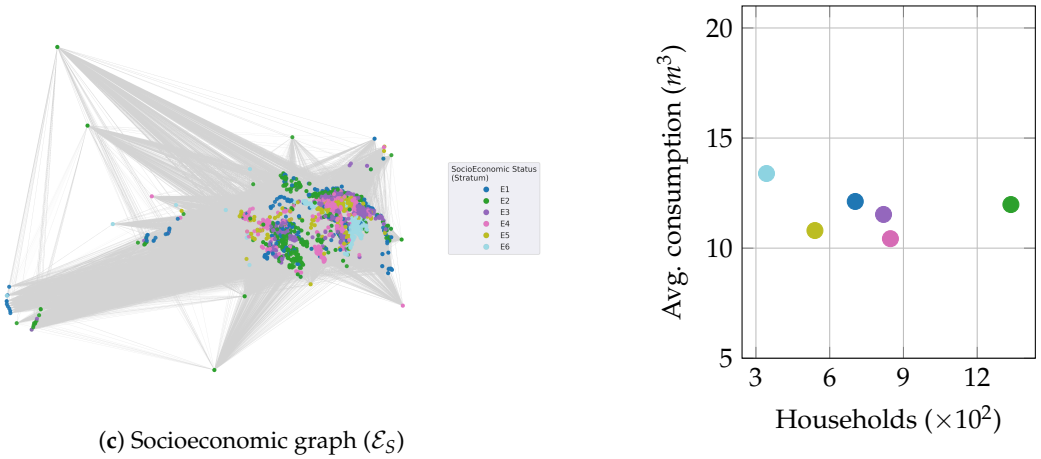
(**a**) Geospatial graph ($\mathcal{E}_G$)



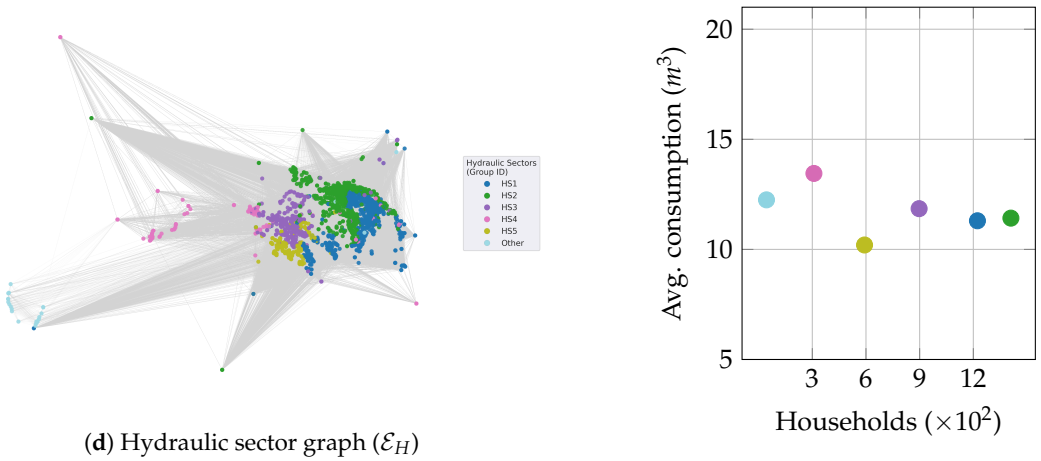
(**b**) Territorial graph ($\mathcal{E}_T$)

Figure 1. *Cont.*

 doi:10.20944/preprints202604.0031.v1



(**c**) Socioeconomic graph ($\mathcal{E}_S$)



(**d**) Hydraulic sector graph ($\mathcal{E}_H$)

Figure 1. Graph structures (left), and average monthly water consumption and number of households (right) according the four predefined household relationships. Since the geospatial graph lacks a grouping variable, the first row only displays the graph structure.

In the training stage, this study considers the following hyperparameter setup: 100 training epochs, 10-epoch-patience early stopping, Adam optimizer at a learning rate of 1×10^{-3}, and weight decay of 5×10^{-4}. Full-batch training is employed to ensure that the complete adjacency structure is available during message passing, which is necessary for the spatial loss term. To account for the stochastic nature of gradient-based optimization, all reported metrics correspond to the average over 10 independent runs with different random seeds; performance was stable across runs, with standard deviations below 0.02 for the Silhouette Score in all configurations. Lastly, a grid search maximizing the Silhouette Score tuned the trade-off parameter in the edge weight computation (see Equation (3)), the latent embedding dimension, and the loss weights, yielding: $\alpha = 0.7$, $d_h = 4$, $\lambda = 0.7$, $\eta = 0.3$, and $\xi = 0.2$.

To assess the individual contribution of each loss term, an ablation study was conducted by systematically removing one component at a time from the composite loss function (Equation (10)) and evaluating the resulting clustering performance under the socioeconomic topology $\mathcal{E}_S$, which yielded the most discriminative results. The full model consistently outperformed all ablated variants, confirming that each term addresses a distinct failure mode: removing $\mathcal{L}_{\text{spatial}}$ led to spatially incoherent assignments, removing $\mathcal{L}_{\text{KL}}$ reduced assignment confidence, removing $\mathcal{L}_{\text{balance}}$ produced degenerate cluster size distributions, and removing $\mathcal{L}_{\text{sep}}$ caused centroid collapse.

4. Results and Discussion

4.1. Quantitative Performance Results

To assess the clustering quality of the proposed Simple GLAC, its performance metrics are compared with those of traditional and graph-based approaches described in Section 3.3. Table 1 summarizes the clustering performance of these methods across four graph topologies. Bold-face

 doi:10.20944/preprints202604.0031.v1

indicates the best-performing method for each graph structure, while the gray-filled cell highlights the overall best approach. The first glance demonstrate inherent limitations of traditional clustering methods (i.e. *K*-means, Hierarchical, and GMM) across different graph settings. Although GMM occasionally achieves competitive Silhouette Scores (for example, $SS = 0.48$ in $\mathcal{E}_G$), this separation is not corroborated by compactness metrics, as indicated by elevated DBS and low CHS values. These findings suggest that clustering based solely on consumption vectors captures statistical separability but does not yield coherent or context-aware clusters.

Table 1. Clustering performance metrics across graph structures and models with respect to Silhouette Score ↑, Calinski-Harabasz Score ↑, Davies-Bouldin Score ↓, and Intercluster Dissimilarity ↑. The arrow points towards better scores. Bold-face indicates the best-performing method for each graph structure. The gray-filled cell highlights the overall best approach.

Graph structure	Method	SS	CHS	DBS	ID
–	**K-means**	0.29	**1908.11**	**1.36**	109.20
	Hierarchical	0.33	1735.70	1.37	109.00
	GMM	**0.48**	728.02	2.36	**141.37**
$\mathcal{E}_G$	**GConv-GMM**	0.30	154.56	2.40	1.08
	Cluster GCN	0.18	66.32	5.49	1.96
	AE-G-GMM	0.01	16.73	17.97	9.96
	GLAC-GCN	0.20	1307.02	1.61	23.28
	Simple GLAC	**0.43**	**124141.04**	**0.53**	**26.17**
$\mathcal{E}_T$	**GConv-GMM**	0.30	323.48	2.21	0.16
	Cluster GCN	**0.79**	3319.40	1.20	1.37
	AE-G-GMM	0.21	236.63	5.38	**64.92**
	GLAC-GCN	0.31	1833.31	1.20	10.22
	Simple GLAC	0.40	**16300.48**	**1.11**	10.17
$\mathcal{E}_S$	**GConv-GMM**	0.40	1840.50	1.24	3.45
	Cluster GCN	0.50	3178.50	0.82	8.42
	AE-G-GMM	–0.01	34.26	15.49	**23.72**
	GLAC-GCN	0.49	2624.86	0.99	3.99
	Simple GLAC	**0.88**	**75477.18**	**0.33**	13.55
$\mathcal{E}_H$	**GConv-GMM**	**0.87**	5321.70	0.51	0.01
	Cluster GCN	0.67	8691.90	0.47	1.88
	AE-G-GMM	–0.02	108.59	12.27	28.56
	GLAC-GCN	0.64	5582.63	0.62	4.96
	Simple GLAC	0.86	**16451.63**	**0.41**	**32.24**

When turning to graph-based approaches, Simple GLAC consistently achieves the most robust multi-metric performance across all graph structures by balancing separation, compactness, and centroid stability. Firstly, within the geospatial graph ($\mathcal{E}_G$), Simple GLAC better balances the metrics, with the lowest DBS and a substantially higher CHS than all competitors. Although its Silhouette Score is marginally lower than that of GMM, the marked improvement in compactness and centroid separation indicates that Simple GLAC prioritizes spatially coherent and homogeneous clusters over isolated pairwise separation. Secondly, in the case of the territorial graph ($\mathcal{E}_T$), although attaining the highest Silhouette Score, ClusterGCN exhibits lower CHS and ID values, suggesting an oversimplified cluster geometry. In contrast, Simple GLAC achieves a significantly higher CHS and a low DBS, indicating well-structured clusters that align with administrative boundaries. Thirdly, the performance differences in the hydraulic graph ($\mathcal{E}_H$) are less pronounced due to its homogeneous, highly connected topology. Although GConv-GMM achieves a high Silhouette Score, its near-zero ID suggests that separation is driven by outliers rather than meaningful centroid spacing. Simple GLAC attains the lowest DBS and the highest ID, demonstrating superior centroid geometry and robustness against diffusion-induced mean collapse. Lastly, the socioeconomic graph ($\mathcal{E}_S$) represents the most discriminative scenario. Simple GLAC outperforms all other methods, achieving the highest Silhouette Score, the lowest Davies–Bouldin Score, and a notably high CHS. The above results demonstrate that Simple

 doi:10.20944/preprints202604.0031.v1

GLAC consistently achieves the most robust multi-metric performance across all graph structures by balancing separation, compactness, and centroid stability. The reported metrics reflect suitable separation and compactness, confirming the effectiveness of clustering-aware edge adaptation and loss design in capturing income-related consumption patterns. Hence, the end-to-end optimization of graph representation learning and clustering objectives enables the generation of interpretable, operationally meaningful consumption profiles in complex urban networks.

Building on the above global metrics, the cluster-level statistics in Table 2 provide structural information about the identified groups for each clustering approach and graph structure. Results demonstrate that Simple GLAC consistently attains robust clustering performance across all graph structures by effectively balancing separation, compactness, and centroid stability, while adapting to the structural constraints imposed by each topology. In graph topologies with strong discriminative capacity, such as geospatial and socioeconomic, the proposed approach produces well-ordered, compact clusters that reflect meaningful consumption patterns, as confirmed by both global validation metrics and low intracluster dispersion. In more constrained settings, such as the territorial graph, Simple GLAC preserves administratively coherent structures without collapsing cluster geometry, while in highly connected and homogeneous topologies, such as the hydraulic graph, it mitigates diffusion-induced homogenization by maintaining centroid separation and isolating consumption extremes. These advantages become more evident when contrasted with baseline approaches. For instance, traditional methods such as K-means and GMM, which ignore relational context, tend to form statistically separable but spatially fragmented clusters that lack coherence with urban structure. GConv-GMM partially incorporates topology but remains limited by its fixed-graph assumptions, resulting in moderate improvements without achieving stable cluster geometry. AE-G-GMM consistently exhibits severe internal dispersion, reflecting the misalignment between reconstruction-driven objectives and clustering quality. ClusterGCN, while achieving low intracluster dispersion, often collapses most nodes into dominant clusters, sacrificing interpretability and intercluster separation. The cluster-level statistics further clarify these behaviors: whereas AE-G-GMM produces highly unstable and diffuse groupings, and ClusterGCN attains compactness through degenerate concentration, Simple GLAC maintains a balanced partitioning, accepting moderate dispersion in minority clusters as a necessary trade-off for preserving meaningful differentiation. Overall, the strong alignment between global metrics and the internal cluster structure demonstrates that the proposed end-to-end framework for Simple Glac not only overcomes the structural limitations of baseline methods but also yields interpretable, operationally relevant consumption profiles in complex urban networks.

Table 2. Cluster-wise statistics for clustering approaches and graph structures.

Graph Structure	Data	GConv-GMM			Cluster GCN			AE-G-GMM			GLAC-GCN			Simple Glac		
	Clusters	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
$\mathcal{E}_G$	Users	1015	2744	831	842	322	3426	2646	1263	681	64	4173	353	60	1913	2617
	Mean (m^3)	11.47	11.55	12.22	11.46	11.57	13.01	11.04	12.25	12.60	10.08	11.54	13.29	10.15	11.37	12.09
	Variance	71.01	58.34	82.67	57.32	63.56	87.93	61.07	69.44	87.07	108.12	60.79	111.70	113.45	65.34	64.12
	IC	21.24	0.57	1.58	12.56	10.52	10.21	4385.72	4989.97	6250.47	139.34	27.51	36.32	**8.54**	**0.26**	**0.25**
$\mathcal{E}_T$	Users	2405	608	1577	3811	43	736	3745	796	49	528	2575	1487	4036	77	477
	Mean (m^3)	11.45	11.47	12.04	11.53	11.86	12.32	10.60	15.71	21.92	11.27	11.66	11.79	11.46	11.95	13.26
	Variance	54.56	71.83	79.86	57.70	139.64	101.76	53.29	92.79	301.66	72.48	61.30	70.59	58.20	57.32	126.84
	IC	0.08	0.02	0.10	0.00	1.72	0.00	3827.12	619.53	21056.55	8.43	28.38	33.02	**0.86**	**1.67**	**24.53**
$\mathcal{E}_S$	Users	645	1630	2315	1332	2748	210	1542	356	2692	784	1936	1870	4171	182	237
	Mean (m^3)	11.73	11.68	12.11	11.29	11.43	13.11	11.34	11.68	14.73	11.21	11.72	11.82	11.53	11.90	13.77
	Variance	66.83	65.26	67.32	60.12	61.30	103.47	62.33	58.65	103.52	52.12	73.79	70.32	61.51	54.49	141.71
	IC	8.95	3.32	2.45	10.63	15.26	6.77	7576.36	4500.43	4823.17	**2.84**	**2.83**	**6.00**	0.53	25.37	20.55
$\mathcal{E}_H$	Users	4475	104	11	991	3537	62	1804	2043	743	936	236	3418	448	4109	33
	Mean (m^3)	11.65	11.97	12.33	11.27	11.75	12.51	10.63	11.08	15.45	11.31	11.62	11.78	11.38	11.73	11.76
	Variance	65.25	81.90	58.52	69.08	63.95	101.67	54.99	65.33	87.08	59.24	70.98	67.41	60.03	65.53	145.42
	IC	8.32	4.01	0.01	**0.15**	**0.03**	**0.14**	3947.84	4681.10	6224.13	0.66	18.47	0.11	27.15	0.37	7.17

4.2. Qualitative Performance Analysis

Figure 2 provides a qualitative comparison of clustering results across graph topologies (columns) and methods (rows), with each cell showing both the full spatial distribution of households and a zoomed view of downtown. The color coding (low, medium, high consumption) allows direct visual assessment of cluster coherence, spatial continuity, and alignment with underlying urban structures. When interpreted jointly with the quantitative results, these maps highlight not only the numerical

 doi:10.20944/preprints202604.0031.v1

performance of each approach but also its effectiveness in translating relational information into meaningful spatial patterns.

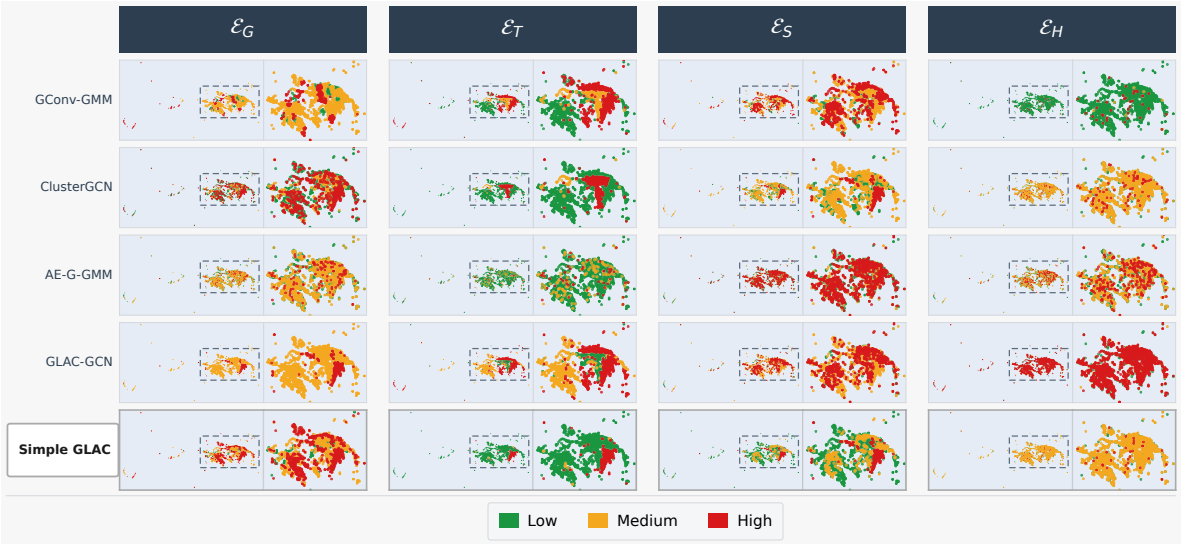


Figure 2. Spatial distribution of clusters across graph topologies (in columns) and graph clustering approaches (in rows). Each cell displays the full spatial distribution of households (left panel) and a zooming view of the central urban core (right panel). Colors denote the estimated consumption clusters: low (■), medium (■), and high (■).

In the $\mathcal{E}_G$ graph, the spatial distributions indicate that Simple GLAC produces a clear, interpretable consumption stratification, characterized by a dominant high-consumption region, a well-defined intermediate band, and a compact low-consumption group. This structure reflects a coherent spatial gradient that aligns with the improved compactness and centroid separation observed in the quantitative metrics. In contrast, baseline methods exhibit notable deficiencies: ClusterGCN collapses most households into a single dominant cluster, yielding limited differentiation; AE-G-GMM produces highly dispersed, spatially incoherent assignments; and GConv-GMM and GLAC-GCN capture intermediate patterns but fail to clearly separate high-consumption areas. These visual patterns are consistent with the cluster-wise statistics in Table 2, where Simple GLAC maintains balanced partitions with low dispersion while preserving meaningful separation. Overall, the geospatial results demonstrate that Simple GLAC effectively leverages local spatial relationships to generate coherent, operationally relevant consumption profiles.

In the case of $\mathcal{E}_T$ topology, where administrative boundaries impose strong diffusion constraints, the clustering results highlight substantial differences in how methods interpret regional structure. Simple GLAC produces the most territorially coherent partition, with clusters that closely follow zone boundaries and capture the underlying consumption gradient without collapsing regional distinctions. This behavior is consistent with its strong balance between compactness and separation. In contrast, GConv-GMM reduces the problem to identifying isolated high-consumption areas, failing to distinguish intermediate regions, and ClusterGCN assigns the majority of households to a single cluster, resulting in an oversimplified, degenerate spatial pattern. AE-G-GMM exhibits fragmented, spatially inconsistent assignments, whereas GLAC-GCN attenuates inter-communal differences through diffusion averaging. The cluster-level statistics confirm that Simple GLAC preserves both structural coherence and differentiation, whereas baseline methods either collapse clusters or inflate dispersion. Thus, the territorial results emphasize the capacity of Simple GLAC to transform the administrative context into actionable consumption segmentation.

The topology based on socioeconomic information ($\mathcal{E}_S$) stands out as the most discriminative and complex, as consumption patterns do not follow a simple monotonic relationship with income. In this setting, Simple GLAC produces a spatially coherent partition that separates high-consumption households, identifies a dominant lower-middle-consumption group, and preserves a transition zone between the two. This organization is consistent with both strong global metrics and well-structured

 doi:10.20944/preprints202604.0031.v1

cluster statistics, indicating that the method successfully captures heterogeneous consumption behavior. In contrast, GLAC-GCN achieves compactness by compressing clusters into a narrow range, effectively eliminating the socioeconomic gradient, while GConv-GMM produces scattered patterns driven by local extremes rather than systematic structure. AE-G-GMM fails to capture any meaningful organization, and ClusterGCN, although capable of reflecting some nuanced behaviors such as elevated consumption at both ends of the income spectrum, lacks overall consistency. These results indicate that Simple GLAC provides the most reliable large-scale socioeconomic stratification, even if certain localized transition effects remain challenging. Therefore, the socioeconomic analysis underscores the importance of aligning clustering objectives with complex real-world patterns.

In the last graph, $\mathcal{E}_H$, the high connectivity and structural homogeneity of the hydraulic sectors impose strong diffusion effects, limiting the separability of consumption patterns across all methods. Within this constrained environment, Simple GLAC maintains the most balanced and stable clustering solution, preserving centroid separation and isolating consumption extremes despite the tendency toward mean convergence. In contrast, GConv-GMM produces a highly degenerate solution dominated by a single cluster, while ClusterGCN achieves near-perfect compactness by concentrating most households into an intermediate group, sacrificing differentiation. GLAC-GCN exhibits clear mean-collapse, with clusters compressed into a narrow range, and AE-G-GMM again produces unstable and dispersed assignments. The cluster-wise statistics confirm that, although all methods are affected by homogenization, Simple GLAC uniquely maintains a meaningful partition by balancing compactness and separation. Therefore, even in structurally constrained settings, the method retains practical value by identifying extreme consumption behaviors within infrastructure-defined regions.

Finally, a consistent pattern emerges across considered graph structures: the effectiveness of clustering methods is fundamentally conditioned by the interplay between graph topology and diffusion dynamics. Baseline approaches fail by ignoring relational structure, collapsing groups under excessive smoothing, or producing unstable representations, leading to either statistically weak or operationally uninterpretable clusters. In contrast, Simple GLAC consistently achieves a balanced trade-off between separation and compactness, adapting to both discriminative and homogeneous settings while preserving meaningful consumption gradients. The alignment between global metrics, cluster-wise statistics, and spatial patterns confirms that its end-to-end optimization framework effectively integrates graph structure and clustering objectives. The above patterns demonstrate that reliable segmentation in urban systems requires not only strong numerical performance but also structural consistency across multiple levels of analysis, a criterion that Simple GLAC satisfies across all evaluated topologies.

5. Conclusions and Future Work

This study introduces and validates Simple GLAC, a novel end-to-end graph-based clustering framework. The proposed approach bridges graph representation learning and sustainable urban water management by providing utility managers with consumption profiles that are not merely statistically compact but operationally actionable across the spatial, administrative, socioeconomic, and infrastructural dimensions that jointly govern urban demand. Simple GLAC includes a graph convolutional encoder with adaptive attention and a topology-agnostic composite loss function. Thanks to the above framework, the proposed approach addresses the limitations of conventional consumption profiling by explicitly modeling the multifaceted relational context of urban water use. The experimental analysis yields several findings of both methodological and practical importance.

Firstly, the quantitative gap between traditional consumption-only methods and graph-based approaches proves that relational context is a key determinant of cluster quality in urban water systems. This finding has direct implications for the development of advanced smart meter analytics platforms. Second, cross-topology analysis shows that no single graph formulation is universally optimal. The socioeconomic topology maximizes behavioral discrimination. The geospatial topology preserves localized demand correlations. The territorial topology aligns cluster boundaries with administrative zones. Third, the

 doi:10.20944/preprints202604.0031.v1

composite loss function of Simple GLAC jointly enforces spatial homogeneity, assignment confidence, size balance, and centroid separation. This produces the most robust multi-metric performance across all four topologies. The outcome shows that single-objective or two-stage methods are less suitable for the heterogeneous, multi-scale structure of urban water networks.

Regarding the considered graph topologies, two structural findings emerge. In the socioeconomic graph, in addition to outperforming baselines, Simple GLAC captures the non-monotonic income–consumption relationship and highlights socioeconomic relationships that aggregate metrics miss. This finding underscores that optimal method selection must be guided not solely by performance indices but by the specific informational requirements of downstream decisions, whether macro-level stratification for tariff policy or fine-grained gradient identification for targeted conservation programs. In the hydraulic graph, the dense intra-sector connectivity induces diffusion-driven convergence toward the citywide mean, constraining all evaluated methods, including Simple GLAC, an inherent property of highly homogeneous graph structures rather than a model-specific deficiency.

Concludingly, Simple GLAC bridges graph representation learning and sustainable urban water management by providing utility managers with consumption profiles that are not merely statistically compact but operationally actionable across the spatial, administrative, socioeconomic, and infrastructural dimensions that jointly govern urban demand. Therefore, the graph topology serves not only as a structural convenience but as a determinant of cluster semantics: identical household data, when encoded geospatially, territorially, socioeconomically, or hydraulically, produces fundamentally different and complementary consumption profiles, each corresponding to a distinct dimension of urban water governance.

Notwithstanding these contributions, three future research directions aim at improving the real-world validity. First, the proposed approach treats the relational graph structure as a fixed, predefined input rather than a learnable component. Integrating adaptive graph construction mechanisms and multi-topology ensemble approaches should allow joint optimization of graph topology and cluster assignments, potentially unraveling relational structures that any single predefined topology misses. Second, while Simple GLAC produces operationally interpretable consumption profiles, the individual contributions of spatial, socioeconomic, and hydraulic factors to each specific cluster assignment remain entangled within the learned latent representation. Developing post-hoc explainability modules applied to the clustering subgraph or attention weight visualization per topology would enable utility managers to understand not only which consumption profile a household belongs to, but which relational factors most strongly condition that assignment [38]. Third, despite being domain-independent by design and requiring only a node-feature matrix and a relational graph as inputs, the Simple GLAC evaluation is conducted on only the water-supply service of a single city due to data availability constraints. Assessing generalizability across different relational structures (e.g., grid topology, tariff zones, building typology) and multiple public utilities (e.g., electricity, natural gas, district heating networks) represents both a methodological validation challenge and a high-priority practical contribution.

Author Contributions: J.A.G and D.C.P. conceptualized the methodology, developed the machine learning methods, and prepared the original draft of the manuscript. J.A.G. was responsible for data curation, investigation, and validation, and contributed to the methodology. D.C.P. contributed to the original draft's preparation. H.F.G. and A.A.O.G. contributed to the conceptualization, development of artificial intelligence methods. J.A.G, and J.G.G.E. took part in writing, reviewing, and editing the original draft. A.A.O.G., J.G.G.E., H.F.G., and D.C.P. were involved in reviewing and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: J. Arias-Garcia is funded by: "Beneficiario proyecto de formación de capital humano de alto nivel" (Conv 22 corte 2).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

 doi:10.20944/preprints202604.0031.v1

Use of Artificial Intelligence: During the preparation of this work the author used Grammarly in order to verify the English grammar. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgments: Under grants provided by the research project 111089082356, funded by MINCIENCIAS.

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

References

1. Liu, H.; Xing, R.; Davies, E.G. Forecasting municipal water demands: Evaluating the impacts of population growth, climate change, and conservation policies on water end-use. *Sustainable Cities and Society* **2025**, *130*, 106581. https://doi.org/https://doi.org/10.1016/j.scs.2025.106581.
2. Yerbury, L.W.; Campello, R.J.; Livingston Jr, G.; Goldsworthy, M.; O'Neil, L. Comparing clustering approaches for smart meter time series: Investigating the influence of dataset properties on performance. *Applied Energy* **2025**, *391*, 125811. https://doi.org/https://doi.org/10.1016/j.apenergy.2025.125811.
3. Jimenez-Castaño, C.; Álvarez Meza, A.; Cárdenas-Peña, D.; Orozco-Gutíerrez, A.; Guerrero-Erazo, J. Kreĭn twin support vector machines for imbalanced data classification. *Pattern Recognition Letters* **2024**, *182*, 39–45. https://doi.org/https://doi.org/10.1016/j.patrec.2024.03.017.
4. Wang, R.; Zhao, X.; Qiu, H.; Cheng, X.; Liu, X. Uncovering urban water consumption patterns through time series clustering and entropy analysis. *Water Research* **2024**, *262*, 122085. https://doi.org/https://doi.org/10.1016/j.watres.2024.122085.
5. Qiao, J.; Shen, K.; Xiao, W.; Tang, J.; Chen, Y.; Xu, J. Integrating Graph Data Models in Advanced Water Resource Management: A New Paradigm for Complex Hydraulic Systems. *Water* **2025**, *17*. https://doi.org/10.3390/w17010003.
6. Khemani, B.; Patil, S.; Kotecha, K.; Tanwar, S. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data* **2024**, *11*, 18. https://doi.org/10.1186/s40537-023-00876-4.
7. Huang, Z.; Tang, Y.; Chen, Y. A graph neural network-based node classification model on class-imbalanced graph data. *Knowledge-Based Systems* **2022**, *244*, 108538. https://doi.org/https://doi.org/10.1016/j.knosys.2022.108538.
8. Zhang, X.; Xie, X.; Kang, Z. Graph Learning for Attributed Graph Clustering. *Mathematics* **2022**, *10*. https://doi.org/10.3390/math10244834.
9. Cárdenas-Peña, D.; Collazos-Huertas, D.; Álvarez Meza, A.; Castellanos-Dominguez, G. Supervised kernel approach for automated learning using General Stochastic Networks. *Engineering Applications of Artificial Intelligence* **2018**, *68*, 10–17. https://doi.org/https://doi.org/10.1016/j.engappai.2017.10.003.
10. Jin, X.; Han, J., K-Means Clustering. In *Encyclopedia of Machine Learning*; Springer US: Boston, MA, 2010; pp. 563–564. https://doi.org/10.1007/978-0-387-30164-8_425.
11. Bar-Joseph, Z.; Gifford, D.K.; Jaakkola, T.S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **2001**, *17*, S22–S29, [https://academic.oup.com/bioinformatics/article-pdf/17/suppl_1/S22/50522365/bioinformatics_17_suppl1_s22.pdf]. https://doi.org/10.1093/bioinformatics/17.suppl_1.S22.
12. Reynolds, D., Gaussian Mixture Models. In *Encyclopedia of Biometrics*; Springer US: Boston, MA, 2009; pp. 659–663. https://doi.org/10.1007/978-0-387-73003-5_196.
13. Bermejo-Martín, G.; Rodríguez-Monroy, C.; Núñez-Guerrero, Y.M. Design Thinking for Urban Water Sustainability in Huelva's Households: Needfinding and Synthesis through Statistic Clustering. *Sustainability* **2020**, *12*. https://doi.org/10.3390/su12219163.
14. Ioannou, A.E.; Creaco, E.F.; Laspidou, C.S. Exploring the Effectiveness of Clustering Algorithms for Capturing Water Consumption Behavior at Household Level. *Sustainability* **2021**, *13*. https://doi.org/10.3390/su13052603.
15. Candelieri, A. Clustering and Support Vector Regression for Water Demand Forecasting and Anomaly Detection. *Water* **2017**, *9*. https://doi.org/10.3390/w9030224.
16. Silva, M.G.; Madeira, S.C.; Henriques, R. Water Consumption Pattern Analysis Using Biclustering: When, Why and How. *Water* **2022**, *14*. https://doi.org/10.3390/w14121954.
17. Padulano, R.; Del Giudice, G. A Mixed Strategy Based on Self-Organizing Map for Water Demand Pattern Profiling of Large-Size Smart Water Grid Data. *Water Resources Management* **2018**, *32*, 3671–3685. https://doi.org/10.1007/s11269-018-2012-7.

 doi:10.20944/preprints202604.0031.v1

18. Tsitsulin, A.; Palowitch, J.; Perozzi, B.; Müller, E. Graph clustering with graph neural networks. *Journal of Machine Learning Research* **2023**, *24*, 1–21.
19. Hou, M.; Xia, F.; Gao, H.; Chen, X.; Chen, H. Urban region profiling with spatio-temporal graph neural networks. *IEEE Transactions on Computational Social Systems* **2022**, *9*, 1736–1747.
20. Jia, Z.; Li, H.; Yan, J.; Sun, J.; Han, C.; Qu, J. Dynamic Graph Convolution-Based Spatio-Temporal Feature Network for Urban Water Demand Forecasting. *Applied Sciences* **2023**, *13*, 10014.
21. Tang, J.; Xia, L.; Huang, C. Explainable Spatio-Temporal Graph Neural Networks. In Proceedings of the Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 2432–2441.
22. Wang, B.; Luo, X.; Zhang, F.; Yuan, B.; Bertozzi, A.L.; Brantingham, P.J. Graph-based deep modeling and real time forecasting of sparse spatio-temporal data. *arXiv preprint arXiv:1804.00684* **2018**.
23. Ma, F.; Liu, F.; Li, W. Jet tagging algorithm of graph network with Haar pooling message passing. *Physical Review D* **2023**, *108*, 072007.
24. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI open* **2020**, *1*, 57–81.
25. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* **2016**.
26. Chiang, W.L.; Liu, X.; Si, S.; Li, Y.; Bengio, S.; Hsieh, C.J. Cluster-GCN: An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks. In Proceedings of the Proceedings of the 25th ACM International Conference on Knowledge Discovery; Data Mining. ACM, jul 2019, KDD '19. https://doi.org/10.1145/3292500.3330925.
27. Wang, C.; Pan, S.; Hu, R.; Long, G.; Jiang, J.; Zhang, C. Attributed Graph Clustering: A Deep Attentional Embedding Approach. In Proceedings of the International Joint Conference on Artificial Intelligence, 2019.
28. Daneshfar, F.; Soleymanbaigi, S.; Yamini, P.; Amini, M.S. A survey on semi-supervised graph clustering. *Engineering Applications of Artificial Intelligence* **2024**, *133*, 108215. https://doi.org/https://doi.org/10.1016/j.engappai.2024.108215.
29. Monti, F.; Boscaini, D.; Masci, J.; Rodolà, E.; Svoboda, J.; Bronstein, M.M. Geometric deep learning on graphs and manifolds using mixture model CNNs, 2016, [arXiv:cs.CV/1611.08402].
30. Wang, C.; Pan, S.; Long, G.; Zhu, X.; Jiang, J. Mgae: Marginalized graph autoencoder for graph clustering. In Proceedings of the Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 889–898.
31. Mrabah, N.; Bouguessa, M.; Touati, M.F.; Ksantini, R. Rethinking graph auto-encoder models for attributed graph clustering. *IEEE Transactions on Knowledge and Data Engineering* **2022**, *35*, 9037–9053.
32. Xu, Y.K.; Huang, D.; Wang, C.D.; Lai, J.H. GLAC-GCN: global and local topology-aware contrastive graph clustering network. *IEEE Transactions on Artificial Intelligence* **2024**.
33. Yang, L.; Yang, R.; Zuo, Z.; Kwan, M.P.; Zhou, S. Graph distance and feature-guided multi-view clustering: A novel method for clustering urban buildings. *Transactions in GIS* **2023**, *27*, 2127–2158.
34. Maaten, L.v.d.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*, 2579–2605.
35. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised deep embedding for clustering analysis. In Proceedings of the International conference on machine learning. PMLR, 2016, pp. 478–487.
36. Petrovic, S. A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In Proceedings of the Proceedings of the 11th Nordic workshop of secure IT systems. Citeseer, 2006, Vol. 2006, pp. 53–64.
37. Wang, X.; Xu, Y. An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. In Proceedings of the IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2019, Vol. 569, p. 052024.
38. Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems* **2019**, *32*.