POLITECNICO DI TORINO
Repository ISTITUZIONALE

A Stability-Aware Consensus Framework for Urban Anomaly Discovery in Multi-Relational POI Graphs

Original
A Stability-Aware Consensus Framework for Urban Anomaly Discovery in Multi-Relational POI Graphs / Vazirov, E., Monaco, S., Apiletti, D.. - In: SMART CITIES. - ISSN 2624-6511. - 9:9(2026). [10.3390/smartcities9090150]

Availability:
This version is available at: 11583/3015704 since: 2026-09-18T21:24:35Z

Publisher:
MDPI

Published
DOI:10.3390/smartcities9090150

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

28 November 2026

smart cities

MDPI

Article

A Stability-Aware Consensus Framework for Urban Anomaly Discovery in Multi-Relational POI Graphs

Etibar Vazirov *, Simone Monaco and Daniele Apiletti *

Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy; simone.monaco@polito.it

* Correspondence: etibar.vazirov@polito.it (E.V.); daniele.apiletti@polito.it (D.A.)

Highlights

What are the main findings?

- A stability-aware consensus framework combines multiple graph autoencoder architectures, anomaly-scoring perspectives, and repeated training runs to provide a more reproducible basis for urban anomaly discovery.
- Evaluation across five geographically distinct cities shows measurable consistency of recurring anomaly patterns across random seeds and reveals partial, method-dependent overlap with representative baseline detectors.

What are the implications of the main findings?

- Stability assessment provides an additional reliability criterion for evaluating unsupervised urban anomaly detection when external ground-truth labels are unavailable.
- The proposed workflow supports reproducible analysis of multi-relational urban POI graphs and can support future research in urban informatics and graph-based urban analytics.

Abstract

Urban anomalies such as rare facilities, unusual spatial configurations, and atypical functional patterns can provide valuable insights into urban dynamics and planning processes. However, graph-based anomaly discovery methods often suffer from instability, producing substantially different results across training runs and making the detected anomalies difficult to reproduce and interpret. In this paper, we use the term *anomaly* to denote automatically detected abnormal urban entities, while the term *urban irregularity* refers to their interpretation within the urban context. We present a consensus-driven framework for stable anomaly discovery using multi-relational point-of-interest (POI) graphs. Urban facilities are represented as nodes connected through multiple spatial and semantic relations, including geographic proximity, shared categories, shared facility types, and region-based associations. Four graph autoencoder architectures (GAE, ResGAE, VGAE, and SAGEAE) are employed to learn node representations, while reconstruction-, cluster-, neighborhood-, and relation-based anomaly scoring strategies are combined with multi-seed stability analysis to identify consensus anomalies. Experiments conducted on five large-scale cities (Baku, Turin, Vienna, Prague, and Kuala Lumpur) show that the proposed framework identifies recurring anomaly patterns across repeated runs and analytical configurations. Comparisons with representative anomaly detectors reveal partial but method-dependent overlap, while cross-city control experiments indicate that a subset of the detected anomalies exhibits non-random semantic and structural correspondence across different urban

Check for updates

Academic Editor: Di Zhu

Received: 15 July 2026
Revised: 29 August 2026
Accepted: 4 September 2026
Published: 9 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

environments. Additional analyses suggest that consensus anomalies are frequently associated with semantically distinctive urban entities, including recreational areas, utility infrastructure, institutional facilities, specialized services, and cultural landmarks. Overall, the results indicate that stability-aware consensus provides a more reproducible and consistent basis for graph-based urban anomaly discovery and supports the interpretation of recurrent anomaly patterns in large-scale urban POI graphs, without requiring ground-truth anomaly labels.

Keywords: urban analytics; point-of-interest graphs; graph neural networks; graph autoencoders; urban irregularities; anomaly detection; embedding stability; consensus analysis; multi-relational graphs; smart cities

1. Introduction

The increasing availability of large-scale geospatial data and open mapping platforms has significantly expanded the scope of data-driven urban analytics. Modern cities can now be represented through rich networks of spatial, semantic, and functional relationships derived from transportation systems, public services, commercial facilities, and human activities. In this context, multi-relational point-of-interest (POI) graphs have emerged as an effective representation for modeling complex urban environments, enabling the integration of geographic proximity, semantic similarity, and functional interactions within a unified analytical framework [1–3].

Graph representation learning has become a fundamental component of contemporary urban computing. By transforming high-dimensional relational data into compact latent representations, graph-based models support a wide range of applications, including region characterization, mobility analysis, facility recommendation, and spatial pattern discovery [4,5]. The ability to simultaneously exploit multiple relation types is particularly valuable in urban settings, where spatial and semantic dependencies jointly shape urban dynamics [6,7]. As a result, graph embeddings increasingly serve as a foundation for uncovering hidden structures and functional patterns in complex city systems.

One emerging application of graph representation learning is the discovery of unusual or irregular urban entities. Detecting facilities that deviate from their surrounding context can provide valuable insights for urban monitoring, infrastructure assessment, land-use analysis, and planning support. Recent advances in graph anomaly detection have produced a variety of embedding-based approaches capable of identifying anomalous nodes and substructures in attributed networks [8,9]. However, most existing studies primarily focus on improving detection performance or developing new architectures, while comparatively less attention has been devoted to the reliability of the resulting anomaly discoveries.

This limitation is particularly important in unsupervised urban environments, where ground-truth anomaly labels are rarely available. The anomalies identified by a graph-based method may vary considerably depending on the selected embedding architecture, random initialization, optimization trajectory, or anomaly scoring strategy. Previous studies have shown that representation learning outcomes can be sensitive to stochastic factors, potentially leading to substantial variability in downstream analyses [10,11]. More broadly, reproducibility has become a growing concern across artificial intelligence and machine learning research, motivating increased attention toward evaluation frameworks that explicitly assess methodological reliability [12]. Despite these concerns, systematic investigations of anomaly stability remain largely absent from the urban graph analytics literature.

From a practical perspective, unstable anomaly detection can significantly reduce the usefulness of urban analytics systems. If repeated executions of the same methodology produce substantially different sets of anomalous facilities, urban planners and decision-makers may struggle to distinguish meaningful urban irregularities from artifacts introduced by stochastic learning processes. Consequently, anomaly discovery frameworks should not only identify unusual entities but also provide evidence that the detected patterns remain consistent under reasonable methodological variations.

Motivated by this challenge, this study investigates anomaly discovery in multi-relational urban POI graphs from a stability and consensus perspective. Rather than proposing a new anomaly detection architecture, we focus on evaluating the reliability of embedding-based urban anomaly discovery under multiple sources of variation. Using graph autoencoder representations, complementary anomaly scoring perspectives, repeated stochastic executions, and multi-city evaluation, we examine whether detected anomalies remain consistent across runs and methods. We further introduce a consensus-driven anomaly selection strategy that prioritizes recurring anomalies identified across multiple embedding architectures, anomaly-scoring perspectives, and random-seed executions, thereby reducing dependence on any single analytical configuration.

The main contributions of this work are threefold. First, we introduce a stability- and consensus-oriented evaluation framework for anomaly discovery in multi-relational urban POI graphs, explicitly quantifying cross-seed consistency and agreement among anomaly detection outcomes. Second, we show how multi-seed stability analysis and consensus-based selection provide a more reproducible and interpretable basis for graph-based anomaly discovery without requiring modifications to the underlying anomaly detection models. Third, we investigate the spatial and semantic characteristics of the resulting consensus anomalies through urban contextualization and evaluate whether independently detected anomalies exhibit measurable, non-random correspondence across different cities through controlled cross-city consistency analysis. Together, these contributions provide a practical framework for more reproducible and interpretable urban POI graph anomaly analysis while explicitly accounting for the absence of external ground-truth anomaly labels.

The remainder of this paper is organized as follows. Section 2 reviews related work on urban graph analytics, graph anomaly detection, and reproducibility-aware learning. Section 3 presents the methodology, including graph construction, embedding learning, anomaly scoring, stability analysis, and consensus discovery. Section 4 reports the experimental protocol and empirical results. Section 5 discusses the main findings, limitations, and future research directions. Finally, Section 6 concludes the paper.

2. Related Work

2.1. Urban POI Graphs and Urban Representation Learning

The increasing availability of large-scale geospatial data has accelerated the adoption of graph-based representations for modeling urban environments. Unlike traditional spatial analysis techniques, graph representations can simultaneously capture geographic proximity, semantic similarity, and functional interactions among urban entities. Recent studies have demonstrated the effectiveness of point-of-interest (POI) graphs and urban knowledge graphs for representing complex city structures and supporting a wide range of downstream urban analytics tasks [1–3].

Graph representation learning has become a fundamental component of modern urban computing. By learning compact latent representations from relational data, graph-based models facilitate region characterization, urban function discovery, mobility analysis, and spatial pattern recognition [4,5]. In particular, multi-relational graph learning approaches

have received growing attention because urban systems naturally involve multiple interacting relationship types. Recent studies have explored relation-aware aggregation mechanisms, heterogeneous graph learning, and dynamic multi-relational representations to better capture the complexity of real-world urban environments [6,7,13,14].

Recent advances further suggest that meaningful urban structures often transcend individual cities. Embedding-based analyses have revealed latent spatial and functional similarities across different urban environments, supporting transferability and cross-city urban understanding [15,16]. These developments motivate the use of graph representations not only for city-specific analysis but also for identifying recurring urban patterns that may generalize across multiple urban contexts.

2.2. Graph Anomaly Detection in Attributed Networks

Graph anomaly detection aims to identify nodes, edges, or substructures that deviate significantly from the dominant patterns of a graph. With the rapid development of graph representation learning, embedding-based anomaly detection has become one of the most active research directions in this field [8,9,17].

Among existing approaches, graph autoencoder architectures have become particularly popular due to their ability to jointly model graph topology and node attributes in an unsupervised manner. Reconstruction-based methods identify anomalies by measuring discrepancies between observed and reconstructed graph information, while probabilistic extensions further account for uncertainty in latent representations [18–20]. More recently, graph anomaly detection has expanded beyond reconstruction-based paradigms. Examples include density-oriented approaches such as Local Outlier Factor (LOF), deep attributed network anomaly detection frameworks such as DOMINANT, and contrastive self-supervised methods such as CoLA [21–24].

Although these approaches have indicated strong anomaly detection capabilities across diverse domains, most studies primarily focus on improving detection performance. Comparatively less attention has been devoted to understanding the consistency and reliability of the anomalies produced by different models and detection paradigms. Consequently, whether different anomaly detection strategies identify similar irregular structures remains an open question, particularly in large-scale urban graph environments.

2.3. Stability, Reproducibility, and Consensus in Unsupervised Learning

The reliability of unsupervised learning outcomes has recently emerged as a growing concern across machine learning research. Multiple studies have shown that learned representations may vary substantially across random initializations, optimization trajectories, and training configurations, potentially affecting downstream analyses and conclusions [10,11]. More broadly, reproducibility has become a central topic in artificial intelligence research, motivating increasing efforts to develop evaluation methodologies that explicitly quantify the robustness and repeatability of learned results [12].

Consensus-based strategies represent one promising direction for improving reliability in anomaly detection. Rather than relying on a single model execution, consensus approaches aggregate evidence from multiple perspectives and prioritize signals that repeatedly emerge across independent analyses. Such mechanisms have been successfully applied in anomaly detection, novelty detection, and ensemble learning settings to reduce uncertainty and improve robustness [25–28]. Similarly, recent work on weakly supervised anomaly detection highlights the importance of validation strategies capable of providing evidence for anomaly relevance when explicit ground-truth labels are unavailable [29].

Despite these developments, the literature still lacks systematic studies that jointly investigate anomaly stability, cross-seed reproducibility, and consensus-driven validation

in multi-relational urban POI graphs. Existing urban graph anomaly detection research typically evaluates anomalies obtained from a single model execution, leaving the robustness of the discovered urban irregularities largely unexplored.

2.4. Research Gap and Positioning of This Work

The reviewed literature demonstrates substantial progress in urban graph representation learning, graph anomaly detection, and reliability-oriented machine learning. However, three important limitations remain. First, most urban anomaly detection studies focus on identifying anomalies rather than assessing whether the detected anomalies remain stable across repeated executions. Second, consensus-based validation mechanisms have rarely been explored in urban graph settings despite their demonstrated value in other anomaly detection domains. Third, relatively little attention has been given to understanding whether the discovered urban irregularities correspond to broader structural patterns that may extend beyond a single city.

This work addresses these limitations by introducing a stability-centered evaluation framework for anomaly discovery in multi-relational urban POI graphs. Rather than proposing a new detection architecture, we investigate the reliability of anomaly discovery across multiple embedding models, anomaly scoring perspectives, and random seeds. Furthermore, we introduce a consensus-driven anomaly selection mechanism and examine the resulting urban irregularities through both quantitative stability analysis and spatial interpretation. In doing so, the study shifts the focus from anomaly detection performance alone toward reproducible and trustworthy urban irregularity discovery.

3. Methodology

3.1. Methodological Overview

Figure 1 presents a conceptual overview of the proposed stability-aware, multi-view graph framework for urban irregularity discovery. Rather than depicting only a sequential experimental workflow, the figure distinguishes the principal methodological components of the framework and their respective roles in robust anomaly discovery.

The framework begins with the construction of city-scale multi-relational POI graphs in which spatial, semantic, and regional relations provide complementary descriptions of urban organization. Heterogeneous latent representations are then learned using four graph autoencoder architectures: GAE, ResGAE, VGAE, and SAGEAE.

The learned representations are examined through four complementary anomaly perspectives: reconstruction error, cluster deviation, neighborhood inconsistency, and relation-conditioned conflict. Each model is trained under repeated random initializations so that variability attributable to stochastic optimization can be measured explicitly rather than being hidden within a single model execution.

Rather than relying on an individual architecture, anomaly score, or stochastic run, the framework integrates repeated detections through stability analysis and consensus-based reliability filtering. Persistent anomaly candidates are subsequently examined through sensitivity analysis, comparison with established and recent anomaly-detection baselines, contextual validation, cross-city correspondence analysis, and urban-semantic interpretation.

The methodological contribution therefore lies at the framework level rather than in proposing a new graph autoencoder architecture or a new ensemble voting rule. Specifically, the proposed methodology integrates heterogeneous graph representation learning, complementary anomaly characterization, repeated-run stability assessment, and consensus-based reliability filtering within a unified multi-city urban irregularity analysis framework.

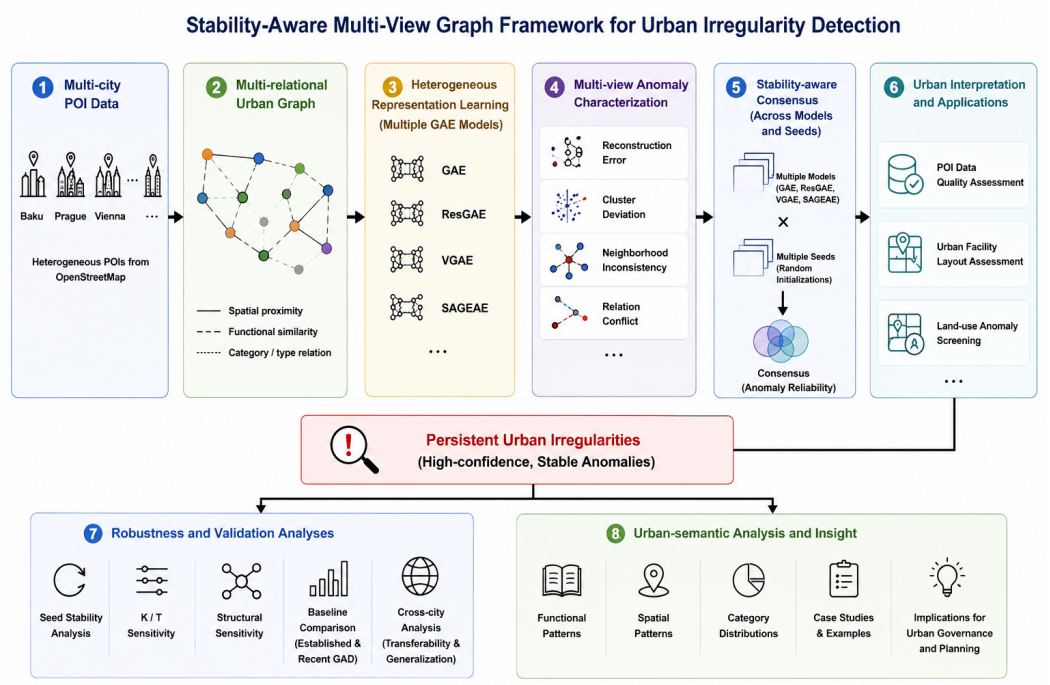


Figure 1. Conceptual overview of the proposed stability-aware multi-view graph framework for urban irregularity detection. Multi-relational urban graphs are analyzed using multiple graph autoencoder architectures and complementary anomaly views. Repeated-seed analysis and consensus filtering are used to identify persistent anomalies, which are subsequently examined through robustness, baseline, cross-city, and urban-semantic analyses and interpreted in the context of urban data quality and planning applications.

3.2. Study Areas and Data Sources

The empirical evaluation is conducted using point-of-interest (POI) data obtained from the Overture Maps ecosystem. POIs constitute a natural unit for urban irregularity discovery because they represent concrete urban facilities, services, and functional locations that collectively characterize the structure and activity patterns of a city. Following recent advances in POI-based urban graph construction and urban knowledge representation, each POI is modeled as an attributed urban entity that can participate in multiple types of spatial and semantic relationships [1,2].

Rather than focusing on a single metropolitan area, this study evaluates anomaly stability and consensus behavior across five cities: Baku (Azerbaijan), Turin (Italy), Vienna (Austria), Prague (Czech Republic), and Kuala Lumpur (Malaysia). These cities were intentionally selected to provide substantial diversity in urban scale, functional composition, and geographic context while maintaining sufficiently rich POI coverage for graph-based analysis. The resulting study areas span multiple regions and exhibit considerable variation in graph size, ranging from approximately 24 thousand POIs in Baku to more than 91 thousand POIs in Kuala Lumpur. Such diversity enables the proposed framework to be evaluated under heterogeneous urban conditions rather than within a single city-specific environment.

The characteristics of the resulting datasets are summarized in Table 1. Considerable variation can be observed in both graph size and feature dimensionality, providing a realistic setting for assessing whether detected urban irregularities remain stable across different urban environments. The use of multiple cities is particularly important in the context of this study, as it allows stability, consensus, and cross-city consistency to be examined beyond a single geographic setting.

Table 1. Characteristics of the selected study cities. Feature dimensionality corresponds to two standardized projected spatial coordinates combined with one-hot-encoded semantic attributes (`basic_category` , `type`, `brand`, and `operating_status`).

City	POIs	Feat. Dim.
Baku	23,656	544
Turin	38,391	1202
Vienna	64,361	1297
Prague	72,379	1166
Kuala Lumpur	91,503	1629

Each POI is represented by a combination of spatial and semantic attributes. The spatial component consists of the two-dimensional coordinates of each POI after projection from WGS84 (EPSG:4326) to a city-specific local UTM coordinate reference system. The projected x- and y-coordinates are standardized independently within each city using StandardScaler. The semantic component is constructed from four categorical attributes: basic_category, type, brand, and operating_status. Missing categorical values are assigned to an Unknown category, after which the categorical attributes are transformed using one-hot encoding.

Consequently, the feature representation of each POI consists of two standardized spatial coordinates followed by the one-hot-encoded semantic attributes. The resulting feature dimensionality F depends on the number of categorical values represented in each city's POI data. This explains the variation in the "Feat. Dim." column of Table 1, from 544 features for Baku to 1629 features for Kuala Lumpur.

After preprocessing, each city is represented by an attributed node feature matrix

$$\mathbf{X} \in \mathbb{R}^{N \times F}, \tag{1}$$

where N denotes the number of POIs and F denotes the resulting feature dimensionality. The regional type, when available, is retained separately and is not included in the input node feature matrix, while the region-cell identifier is used for relational graph construction rather than as an input node feature. The resulting feature matrix provides the attribute component of the urban graph representation, while relational dependencies between POIs are modeled through the multi-relational graph construction process described next.

3.3. Multi-Relational Urban Graph Construction

Urban irregularities rarely emerge from a single source of information. A facility may appear unusual because of its geographic surroundings, its semantic context, or inconsistencies between multiple relational dimensions. To capture these complementary perspectives, each city is represented as a multi-relational graph

$$G = (V, \mathcal{E}, \mathbf{X}), \tag{2}$$

where V denotes the set of POI nodes, $\mathbf{X}$ represents the node feature matrix, and $\mathcal{E} = E^{(1)}, E^{(2)}, \ldots, E^{(R)}$ denotes the collection of relation-specific edge sets.

Three categories of relations are incorporated into the graph construction process. Spatial relations connect geographically nearby facilities and capture local urban continuity. Semantic relations connect facilities sharing similar functional characteristics, enabling the graph to represent similarities in urban purpose and activity. Regional relations connect facilities belonging to the same spatial partition and capture medium-scale organizational structures that are not necessarily reflected through local proximity alone. Together, these

relation types provide complementary views of urban organization and allow irregularities to be analyzed from multiple relational perspectives [6,7].

In the present study, five relation types are constructed: two spatial relations, two semantic relations, and one regional relation. The spatial relations connect POIs within 75 m (geo-close) and 250 m (geo-medium), respectively. These thresholds are applied after projecting each city's POIs to a local metric coordinate reference system. The two scales are intended to capture complementary spatial contexts: the 75 m relation represents immediate local proximity, whereas the 250 m relation captures a broader neighborhood context. Semantic relations connect POIs sharing the same basic category or the same facility type, while the regional relation connects POIs assigned to the same 200 m $\times$ 200 m grid cell. All relation types are represented as unweighted binary edges. To limit excessive connectivity in dense spatial neighborhoods, the graph-construction procedure caps the number of geo-close and geo-medium neighbors per node at 30 and 50, respectively. For large semantic or regional groups, local group-neighbor connections are used rather than unrestricted full pairwise connectivity.

The 75 m and 250 m spatial thresholds constitute the default graph configuration used in the main experiments. To assess whether the structural characteristics of the constructed graphs depend strongly on these particular choices, an additional lightweight structural sensitivity analysis varies the close-distance threshold over 50, 75, and 100 m and the medium-distance threshold over 200, 250, and 300 m. The corresponding results are reported in the sensitivity analysis, providing an empirical robustness check around the default spatial configuration.

For each relation type r, an adjacency matrix $\mathbf{A}^{(r)}$ is constructed. To obtain a unified topology for representation learning, the relation-specific adjacency matrices are merged into a single graph while preserving whether at least one relation exists between two POIs:

$$\mathbf{A} = \mathbb{I}\left(\sum_{r=1}^{R} \mathbf{A}^{(r)} > 0\right), \tag{3}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

The resulting unified adjacency matrix therefore records whether at least one of the defined relations connects a pair of POIs, while duplicate node pairs arising from multiple relation types are represented by a single edge in the training topology. Importantly, the union operation is used only to construct the topology supplied to the graph autoencoder encoders; the individual relation-specific adjacency matrices and edge sets are retained separately for relation-aware downstream analysis.

The resulting graphs contain tens of thousands of nodes and hundreds of thousands to several million edges depending on the city. Such scale substantially exceeds that of many commonly used graph anomaly detection benchmarks and provides a realistic environment for evaluating the robustness of urban irregularity discovery under heterogeneous graph structures.

All four graph autoencoder architectures learn a single node-level latent representation from the unified adjacency matrix and the node feature matrix. Relation-specific latent embeddings are not learned, and no separate encoder parameters are maintained for different relation types. Instead, the original relation-specific edge sets are retained alongside the unified graph and are used during relation-aware anomaly scoring. In particular, relational inconsistency is evaluated by comparing latent representations of nodes connected under the retained relation-specific edge sets. Thus, the unified graph determines the topology used for representation learning, whereas the relation-specific graphs provide distinct relational contexts for subsequent anomaly assessment. This separation avoids implying that

relation identity is explicitly encoded by the autoencoder while still allowing relation-level structural inconsistencies to contribute to irregularity detection.

The resulting multi-relational graphs form the structural foundation for representation learning. The next subsection describes how graph autoencoder-based embedding models transform these attributed urban graphs into latent representations suitable for anomaly scoring, stability assessment, and consensus-driven irregularity discovery.

3.4. Graph Embedding Learning

The multi-relational graphs described in the previous subsection provide a rich representation of urban structure. However, anomaly discovery cannot be performed directly on the high-dimensional graph representation. Instead, it is first necessary to transform the attributed graph into a compact latent space capable of capturing both structural and semantic dependencies among POIs.

To this end, node embeddings are learned using graph autoencoder-based representation learning models. Graph autoencoders have become one of the most widely adopted approaches for unsupervised graph representation learning because they jointly exploit node attributes and graph topology while requiring no labeled data [18]. In the context of this study, the objective of embedding learning is not anomaly detection itself, but rather the construction of latent representations that can subsequently be analyzed from multiple anomaly-scoring perspectives.

Given a graph adjacency matrix $\mathbf{A}$ and node feature matrix $\mathbf{X}$, an encoder maps each node into a latent representation:

$$\mathbf{Z} = f_\theta(\mathbf{X}, \mathbf{A}) \tag{4}$$

where $\mathbf{Z} \in \mathbb{R}^{N \times d}$ denotes the learned embedding matrix and d represents the latent dimensionality.

To reduce architecture-specific bias, four complementary embedding models are considered: Graph Autoencoder (GAE), Residual Graph Autoencoder (ResGAE), Variational Graph Autoencoder (VGAE), and GraphSAGE-based Autoencoder (SAGEAE). Together, these models provide deterministic, residual-enhanced, probabilistic, and inductive representation-learning perspectives, enabling the stability of anomaly discovery to be evaluated across different embedding paradigms rather than within a single architecture.

3.4.1. Graph Autoencoder (GAE)

The baseline Graph Autoencoder employs graph convolutional layers to aggregate neighborhood information and generate latent node representations. Each layer is computed as

$$\mathbf{H}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right) \tag{5}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with self-loops, $\tilde{\mathbf{D}}$ is the corresponding degree matrix, $\mathbf{W}^{(l)}$ denotes trainable parameters, and $\sigma(\cdot)$ is a nonlinear activation function.

The model is optimized through joint reconstruction of graph structure and node attributes [18].

3.4.2. Residual Graph Autoencoder (ResGAE)

Although graph convolutional architectures effectively capture neighborhood information, deeper models may suffer from representation degradation and over-smoothing. To mitigate this effect, a residual pathway is incorporated into the latent representation [30]. The resulting embedding is defined as

$$\mathbf{Z}_{\text{Res}} = f_\theta^{\text{GCN}}(\mathbf{X}, \mathbf{A}) + g_\phi(\mathbf{X}) \tag{6}$$

where $g_\phi(\cdot)$ denotes a multilayer perceptron operating directly on node attributes. This formulation preserves attribute-specific information that may otherwise be weakened through repeated neighborhood aggregation.

3.4.3. Variational Graph Autoencoder (VGAE)

To account for uncertainty in latent representations, a probabilistic embedding formulation is additionally considered through the Variational Graph Autoencoder (VGAE) [31]. Instead of learning a deterministic embedding, each node is represented by a Gaussian latent distribution

$$q(\mathbf{z}_i \mid \mathbf{X}, \mathbf{A}) = \mathcal{N}\Big(\boldsymbol{\mu}_i, \mathrm{diag}(\sigma_i^2)\Big) \tag{7}$$

where $\boldsymbol{\mu}_i$ and σ_i are learned by graph convolutional encoders. The probabilistic formulation enables the model to represent uncertainty and variation in complex graph structures.

3.4.4. GraphSAGE Autoencoder (SAGEAE)

Finally, an autoencoder variant based on GraphSAGE aggregation is included to provide a complementary neighborhood-aggregation mechanism [32]. Unlike standard graph convolutional architectures, GraphSAGE updates node representations by aggregating information from neighboring nodes:

$$\mathbf{h}_v^{(l+1)} = \sigma\Big(\mathbf{W}^{(l)} \cdot \mathrm{AGG}\Big(\Big\{\mathbf{h}_u^{(l)} : u \in \mathcal{N}(v)\Big\}\Big)\Big) \tag{8}$$

where $\mathcal{N}(v)$ denotes the neighborhood of node v and $\mathrm{AGG}(\cdot)$ represents an aggregation operator. GraphSAGE has indicated more scalability on large graphs and provides a complementary embedding mechanism compared with conventional graph convolutional models.

The resulting embeddings constitute the foundation for anomaly discovery. However, embeddings alone do not determine which facilities should be considered unusual. The next subsection therefore introduces multiple anomaly-scoring perspectives designed to identify urban irregularities from complementary structural and semantic viewpoints.

3.5. Multi-Seed Training Strategy

A recurring challenge in unsupervised graph representation learning is the variability of learned embeddings across repeated executions of the same model. Even when the graph structure, node features, and hyperparameter settings remain unchanged, differences in random initialization and stochastic optimization may lead to different latent representations and consequently different downstream outcomes. Previous studies have reported that graph embeddings can exhibit considerable variation across runs, raising concerns regarding reproducibility and reliability in graph-based analysis [10,11].

This issue is particularly relevant for anomaly discovery. Since anomaly rankings are derived from learned embeddings, a facility identified as anomalous in one execution may not necessarily appear among the most anomalous instances in another execution of the same model. As a result, anomaly rankings obtained from a single run may partially reflect stochastic training effects rather than persistent structural irregularities.

To explicitly account for this source of uncertainty, each embedding architecture is trained multiple times using different random seeds. Let

$$S = s_1, s_2, \ldots, s_M, \tag{9}$$

denote the set of random seeds, where M represents the total number of independent executions. For each seed, the model is trained using the same graph, feature matrix, architecture, and hyperparameter configuration. Consequently, any variation observed in

the resulting anomaly rankings can be attributed primarily to stochastic training effects rather than differences in data or model design.

Rather than treating a single execution as definitive, the proposed framework evaluates anomaly behavior across all runs. This design enables the robustness of detected irregularities to be examined directly and aligns with recent efforts emphasizing reproducibility and reliability in machine learning research [12].

The repeated executions generate multiple anomaly rankings for each embedding architecture. These rankings are subsequently analyzed through several complementary anomaly-scoring perspectives, allowing the consistency of detected irregularities to be assessed before stability and consensus evaluation are performed.

3.6. Multi-Perspective Anomaly Scoring

The embedding models described in the previous subsections generate latent representations of urban POIs. However, embeddings alone do not determine whether a facility should be considered irregular. Urban irregularities may manifest in different forms and at different structural scales. A facility may appear unusual because it is poorly reconstructed from its latent representation, deviates from the characteristics of its broader cluster, differs substantially from its immediate neighborhood, or exhibits inconsistent behavior across different relational contexts. Consequently, relying on a single anomaly definition may provide only a partial view of urban irregularity.

To capture these complementary perspectives, anomaly scores are computed using four distinct criteria. Each criterion emphasizes a different manifestation of irregular behavior in the learned embedding space. The scoring perspectives used in this study are summarized in Table 2.

Table 2. Anomaly-scoring perspectives considered in the proposed framework.

Score	Intuition	Captures
Reconstruction	Poor reconstruction quality	Structural deviation
Neighborhood	Mismatch with neighbors	Local inconsistency
Cluster	Distance from cluster pattern	Community deviation
Conflict	Disagreement across relations	Multi-relational irregularity

3.6.1. Reconstruction-Based Irregularity

The first perspective evaluates how accurately a node can be reconstructed from the learned latent representation. Reconstruction-based anomaly detection has been widely adopted in graph representation learning because structurally unusual nodes tend to exhibit larger reconstruction errors than regular instances [22]. For each node, the reconstruction score is defined as

$$A_{\text{rec}}(i) = |\mathbf{x}_i - \hat{\mathbf{x}}_i|_2, \tag{10}$$

where $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ denote the original and reconstructed feature vectors, respectively.

3.6.2. Neighborhood Inconsistency

Urban facilities are typically influenced by their surrounding environment. Consequently, a node whose embedding differs substantially from those of its local neighbors may indicate an unusual local configuration. Inspired by neighborhood-aware anomaly detection approaches [33], the neighborhood inconsistency score is computed as

$$A_{\text{nbr}}(i) = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} |\mathbf{z}_i - \mathbf{z}_j|_2, \tag{11}$$

where $\mathcal{N}(i)$ denotes the neighborhood of node i.

3.6.3. Cluster Deviation

Some irregularities are not apparent at the local level but emerge when a node is compared with the broader group to which it belongs. Following cluster-oriented anomaly analysis [34], each node is compared with the centroid of its assigned embedding cluster:

$$A_{\text{clust}}(i) = \left| \mathbf{z}_i - \mathbf{c}_{k(i)} \right|_2, \tag{12}$$

where $\mathbf{c}_{k(i)}$ denotes the centroid of the cluster containing node i.

3.6.4. Relational Conflict

Multi-relational urban graphs provide complementary relational views of the same facility. Although the graph autoencoders learn a single node-level latent representation from the unified adjacency matrix, the original relation-specific edge sets are retained and can be used to evaluate whether a node behaves consistently across different relational contexts. In the reported experiments, the conflict score compares two complementary views: immediate geographic proximity, represented by the geo-close relation, and semantic similarity, represented by the same-basic-category relation.

For node i, the mean latent-space cosine similarity to its geo-close neighbors is defined as

$$s_{\text{geo}}(i) = \frac{1}{|\mathcal{N}_{\text{geo}}(i)|} \sum_{j \in \mathcal{N}_{\text{geo}}(i)} \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}, \tag{13}$$

where $\mathcal{N}_{\text{geo}}(i)$ denotes the neighbors of node i under the geo-close relation. Similarly, the mean similarity to semantically related POIs is

$$s_{\text{sem}}(i) = \frac{1}{|\mathcal{N}_{\text{sem}}(i)|} \sum_{j \in \mathcal{N}_{\text{sem}}(i)} \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}, \tag{14}$$

where $\mathcal{N}_{\text{sem}}(i)$ denotes the neighbors of node i under the same-basic-category relation.

$$A_{\text{conf}}(i) = \left| s_{\text{geo}}(i) - s_{\text{sem}}(i) \right|, \tag{15}$$

A high conflict score therefore indicates that the node exhibits substantially different latent-space consistency under geographic and semantic relations. For example, a POI may be highly similar to facilities of the same semantic category while being poorly aligned with the POIs located in its immediate geographic surroundings, or vice versa. The score is set to zero when either the geo-close or same-basic-category neighborhood is empty, because the relational comparison cannot be computed in that case.

Importantly, $\mathbf{z}_i$ denotes the single shared embedding learned from the unified graph; no relation-specific latent embeddings or relation-specific encoder parameters are learned. The relation identity enters the conflict score only through the corresponding relation-specific neighborhood sets. This formulation is therefore consistent with the representation-learning procedure described in Section 3.3 while still allowing disagreement between geographic and semantic contexts to contribute to anomaly discovery.

This cosine-similarity-based formulation is the relational conflict score used in all experiments reported in this study.

3.7. Stability Analysis

The anomaly-scoring perspectives introduced in the previous subsection generate multiple rankings of potentially irregular urban facilities. However, the reliability of these rankings cannot be assumed a priori. Since graph embeddings are obtained through stochastic optimization, repeated executions of the same model may produce different

latent representations and consequently different anomaly rankings. Variability across graph embedding runs has previously been observed in representation learning studies, highlighting the need for explicit consistency assessment before drawing conclusions from detected anomalies [10].

The challenge becomes particularly important in unsupervised anomaly discovery, where the absence of ground-truth labels makes validation inherently difficult. More broadly, reproducibility and repeatability have emerged as critical concerns in machine learning research, especially when model outputs are used to support downstream decision-making [11,12]. Consequently, evaluating whether detected anomalies persist across repeated executions becomes a prerequisite for trustworthy urban irregularity analysis.

To assess stability, each embedding architecture is trained multiple times using different random seeds. For every execution, nodes are ranked according to their anomaly scores. Rather than relying on score thresholds, which may vary substantially across models and scoring perspectives, the analysis focuses on the highest-ranked anomalies. This Top-K strategy emphasizes the most prominent irregularities while ensuring comparability across independent runs.

Let

$$A^{(s_m)} = a_1, a_2, \ldots, a_K \tag{16}$$

denote the Top-K anomaly set obtained from the execution associated with random seed s_m, where K represents the number of retained anomalies.

A stable anomaly detection process should repeatedly identify similar anomalies across independent executions. To quantify this behavior, pairwise agreement between Top-K anomaly sets is first measured through their intersection size:

$$I_{ij} = \left| A^{(s_i)} \cap A^{(s_j)} \right| \tag{17}$$

where larger values indicate stronger agreement between the anomaly rankings generated by seeds s_i and s_j.

Although intersection size provides an intuitive measure of overlap, its interpretation depends on the selected value of K. A normalized measure is therefore additionally required. Following established practices in consistency and agreement analysis, similarity between two anomaly rankings is evaluated using the Jaccard coefficient:

$$J_{ij} = \frac{\left| A^{(s_i)} \cap A^{(s_j)} \right|}{\left| A^{(s_i)} \cup A^{(s_j)} \right|} \tag{18}$$

where $J_{ij} \in [0, 1]$, and larger values indicate greater stability across repeated executions [27].

The resulting overlap statistics provide a direct measure of ranking robustness. High agreement suggests that the detected irregularities correspond to persistent structural patterns rather than random optimization artifacts. Conversely, low agreement indicates that anomaly rankings may be sensitive to stochastic effects, a phenomenon that has also been documented in broader deep-learning stability studies [35].

The complete stability evaluation procedure is illustrated in Figure 2. For each embedding architecture and anomaly-scoring perspective, rankings obtained from multiple seeds are compared through Top-K overlap analysis before recurrent anomalies are identified.

The next subsection builds on this stability analysis by introducing the consensus-based selection strategy used to identify recurrent anomaly candidates.

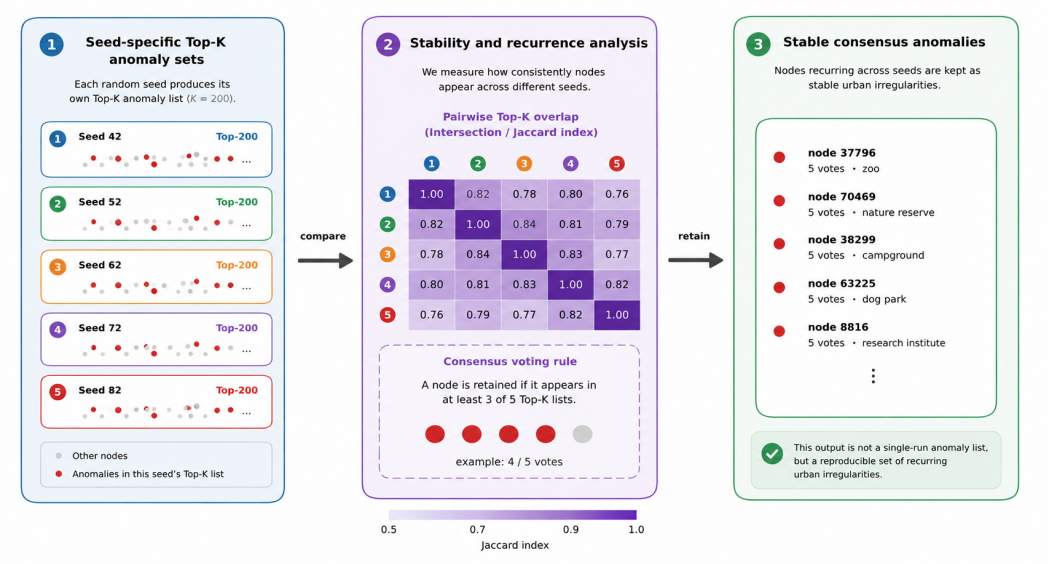


Figure 2. Stability evaluation and consensus discovery workflow. Independent executions generate anomaly rankings that are subsequently compared through Top-K overlap analysis before consensus anomalies are identified. Colors distinguish the five random seeds, red and gray markers denote Top-K anomaly candidates and other nodes, respectively, and the heatmap color intensity represents the Jaccard similarity between seed-specific Top-K sets.

3.8. Consensus-Based Urban Irregularity Discovery

The stability analysis presented in the previous subsection quantifies the consistency of anomaly rankings across repeated executions. However, stability alone does not determine which anomalies should ultimately be retained for interpretation. A node may appear in several rankings while remaining absent from many others. Consequently, an additional selection mechanism is required to identify the most reliable anomalies.

Rather than relying on a single anomaly ranking, the proposed framework adopts a consensus-oriented strategy that aggregates evidence across multiple embedding architectures, anomaly-scoring perspectives, and random-seed executions. The underlying intuition is that anomalies recurring across multiple independent views of the urban graph provide stronger evidence than those detected under only a single configuration. Similar agreement-based principles have been explored in ensemble anomaly detection settings to improve robustness and reduce sensitivity to individual model behavior [36].

Let $\mathcal{R}$ denote the collection of anomaly rankings generated across embedding architectures, anomaly-scoring perspectives, and random-seed executions.

Across the complete experimental design, four embedding architectures (GAE, ResGAE, VGAE, and SAGEAE), four anomaly-scoring perspectives (reconstruction, cluster, neighborhood consistency, and relational conflict), and five random-seed executions produce 80 anomaly rankings per city. These rankings support two complementary consensus analyses: seed-level consensus evaluates recurrence across the five seeds within each fixed model–score configuration, whereas cross-configuration consensus aggregates evidence across the broader collection of model, score, and seed combinations. For a node i, a consensus vote is defined as

$$V(i) = \sum_{r \in \mathcal{R}} \mathbf{1}(i \in A_r), \tag{19}$$

where A_r denotes the Top-K anomaly set associated with ranking r, and $\mathbf{1}(\cdot)$ is the indicator function.

The main experiments use $K = 200$ as the reference candidate-set size, such that each ranking contributes its 200 highest-ranked anomalies to the voting procedure. Because the five city graphs differ substantially in size, the robustness of this fixed-size definition is additionally examined using alternative values of K and proportional candidate sets defined relative to the total number of nodes. This complementary analysis distinguishes sensitivity to an absolute candidate-set size from sensitivity to the fraction of the urban graph designated as anomalous.

Nodes receiving higher vote counts are repeatedly identified as anomalies across multiple independent analyses. A consensus anomaly set is therefore defined as

$$C = \{i : V(i) \geq T\}, \tag{20}$$

where T denotes the minimum agreement threshold required for retention.

The reference configuration uses $T = 3$, meaning that a node must occur in at least three contributing anomaly sets to be retained as a consensus anomaly. Rather than treating this value as an intrinsically optimal threshold, we use it as a moderate agreement requirement and explicitly evaluate alternative consensus thresholds in the sensitivity analysis. The joint sensitivity of K and T is examined to determine how candidate-set breadth and the required level of agreement affect the size and composition of the resulting consensus set. The reference values $K = 200$ and $T = 3$ are retained because they provide a practical balance between overly permissive consensus sets and overly restrictive selections, while the principal conclusions are evaluated for robustness to nearby parameter choices.

In the absence of ground-truth anomaly labels, repeated agreement across different embedding architectures, anomaly-scoring perspectives, and random-seed executions is used as a stability-based source of evidence rather than as a substitute for external ground truth. Nodes that repeatedly reappear under heterogeneous configurations are less dependent on a particular model, score, or stochastic initialization. The consensus mechanism therefore targets reproducibility of anomaly discovery, while complementary baseline, contextual, and domain-oriented validation analyses are used subsequently to assess whether the retained candidates are also meaningful beyond internal agreement.

Furthermore, different anomaly-scoring perspectives emphasize different manifestations of urban irregularity. Reconstruction-, neighborhood-, cluster-, and conflict-based scores may highlight partially overlapping subsets of nodes. Retaining only anomalies that repeatedly emerge across these complementary perspectives reduces dependence on any single anomaly definition and yields a more robust characterization of unusual urban structures [23].

Because the reference experiments use five random seeds, we additionally examine whether the observed consensus behavior depends strongly on this seed count. Seed-count sensitivity is evaluated by progressively varying the number of contributing random-seed executions and measuring the resulting stability of the consensus selections. This analysis complements the K–T sensitivity analysis by testing robustness with respect to the amount of stochastic replication used to construct the consensus.

Under the reference configuration, the resulting consensus candidates constitute the set retained for subsequent urban interpretation and validation. Importantly, consensus membership is interpreted as evidence of reproducible detection across heterogeneous model and scoring configurations, rather than as definitive proof that a POI is erroneous or anomalous in an external ground-truth sense. This distinction is particularly important in the present unsupervised setting and motivates the additional validation analyses reported in the experimental section.

To further assess the behavior of the proposed framework, the following subsection compares the resulting anomaly rankings against representative baseline methods commonly used in graph and non-graph anomaly detection.

3.9. Baseline Methods

The consensus anomalies identified by the proposed framework represent urban irregularities that persist across multiple embedding architectures, anomaly-scoring perspectives, and random-seed executions. While this consensus-oriented strategy emphasizes reliability and recurrence, it is also informative to examine how the resulting irregularities relate to those highlighted by established anomaly detection methods.

The objective of this comparison is not to demonstrate superiority over existing detectors. Different anomaly detection paradigms are designed to capture different manifestations of irregular behavior and therefore may emphasize substantially different subsets of nodes. Instead, baseline methods are used as external reference points that provide additional context for interpreting the irregularities identified by the proposed framework.

Three representative anomaly detection approaches are considered. First, Local Outlier Factor (LOF) is included as a classical density-based method that identifies observations located in sparse regions relative to their local neighborhoods [21]. Second, DOMINANT is adopted as a reconstruction-based graph anomaly detector that jointly models structural and attribute reconstruction errors [22]. Third, CoLA is employed as a contrastive graph anomaly detection approach that identifies irregular nodes through self-supervised local-context discrimination [23].

These baselines were intentionally selected because they represent complementary anomaly detection principles. LOF emphasizes local density deviations, DOMINANT focuses on reconstruction inconsistencies, and CoLA captures contrastive contextual irregularities. Together, they provide diverse perspectives that complement the consensus-oriented framework proposed in this study.

For each baseline method, anomaly scores are computed for all nodes and converted into Top-K anomaly rankings using the same ranking protocol employed throughout the proposed framework. The resulting rankings are subsequently compared with the consensus-supported urban irregularities identified in Section 3.8.

It is important to emphasize that the baseline rankings are not incorporated into the consensus-voting process. Instead, they serve exclusively as external reference points for assessing whether the recurring irregularities identified by the proposed framework are also highlighted by widely used anomaly detection paradigms.

Because unsupervised anomaly discovery lacks ground-truth labels, overlap with baseline detectors alone cannot fully validate the detected irregularities. To further support the interpretation of the discovered anomalies, the following subsection examines their spatial and semantic characteristics within the urban context.

3.10. Urban Contextualization of Consensus Anomalies

The previous stages of the framework focus on identifying statistically persistent irregularities through multi-seed training, multiple embedding architectures, and complementary anomaly scoring perspectives. While these procedures improve the reliability of anomaly discovery, they do not directly explain the urban significance of the detected patterns. Consequently, a final contextualization stage is introduced to examine how consensus-supported anomalies manifest within the geographic and functional structure of the city.

Rather than evaluating anomalies solely in the latent embedding space, the detected consensus anomaly set is projected back into the original urban environment. This enables

the analysis of anomaly patterns from two complementary perspectives: spatial distribution and functional composition. Recent studies have shown that graph-based urban representations can capture meaningful spatial and semantic structures, making geographic interpretation an important component of urban analytics workflows [15,37].

To characterize the functional composition of the consensus anomaly set, category-level frequencies are computed. For a given category (c), the frequency within the consensus anomaly set (C) is defined as

$$f_c = \sum_{i \in C} \mathbf{I}(y_i = c) \tag{21}$$

where y_i denotes the category associated with anomaly node i, and $\mathbf{I}(\cdot)$ is the indicator function. This aggregation provides a compact summary of the urban functions that most frequently appear among consensus-supported irregularities.

In parallel, consensus anomalies are visualized in geographic space to examine their spatial organization and their relationship to surrounding urban structures. Representative examples are subsequently inspected in the context of their local neighborhoods, allowing the analysis to move beyond purely numerical anomaly scores and towards interpretable urban patterns. Similar contextual interpretation strategies have been shown to improve the understanding of graph-derived urban representations and functional urban structures [38].

The resulting spatial visualizations and category-level summaries provide an interpretable urban perspective on the detected anomalies. Building upon this analysis, the next stage investigates whether similar anomaly patterns emerge consistently across geographically distinct cities.

3.11. Cross-City Anomaly Consistency Analysis

While the previous analyses focus on identifying stable consensus anomalies within individual cities, an additional question is whether such anomalies exhibit similar characteristics across different urban environments. To investigate this aspect, we performed a cross-city anomaly consistency analysis. Rather than transferring anomaly labels across cities, the objective is to examine whether independently detected anomalies share comparable semantic and structural properties across geographically distinct urban regions [39].

For each source city, consensus anomalies obtained from the multi-seed consensus process were selected as query nodes. Given a source anomaly node i, candidate nodes from a target city were ranked using a combined similarity score:

$$S(i,j) = w_{\text{emb}} S_{\text{emb}}(i,j) + w_{\text{nbr}} S_{\text{nbr}}(i,j) + w_{\text{sem}} S_{\text{sem}}(i,j) + w_{\text{prof}} S_{\text{prof}}(i,j) \tag{22}$$

where S_{emb} denotes latent embedding similarity, S_{nbr} measures similarity between neighborhood category distributions, S_{sem} captures semantic similarity between POI categories, and S_{prof} compares detailed local category profiles. Multiple weighting schemes were evaluated to assess the relative contribution of structural and semantic information during the matching process.

For each source anomaly, the Top-N highest-ranked nodes from the target city were retrieved. A successful correspondence was recorded if at least one retrieved node belonged to the consensus anomaly set of the target city. The resulting Hit@N score was computed as

$$\text{Hit@}N = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\big(\exists j \in \text{Top}N(i) : j \in \mathcal{A}_{\text{target}}\big), \tag{23}$$

where M denotes the number of source-city consensus anomalies, $\mathcal{A}_{\text{target}}$ represents the consensus anomaly set of the target city, and $\mathbb{I}(\cdot)$ is the indicator function.

The analysis evaluates whether independently detected consensus anomalies exhibit non-random semantic or structural correspondence across different urban environments. It is not intended to establish direct anomaly-label transfer or universal cross-city generalization. Instead, the objective is to determine whether selected anomaly candidates share measurable characteristics across cities and to assess the extent to which such correspondence can be explained by semantic compatibility, structural similarity, or random retrieval. Such cross-city comparison remains relevant in urban graph representation learning, where learned representations can capture functional similarities across heterogeneous urban environments [5].

4. Experimental Protocol and Empirical Results

4.1. Experimental Configuration

Experiments were conducted on five urban POI graphs representing Baku, Prague, Vienna, Turin, and Kuala Lumpur. For each city, graph autoencoder models were trained independently using five random seeds (42, 52, 62, 72, and 82) to assess the stability of anomaly detection under different initialization conditions. The evaluated models included GAE and VGAE.

All four graph autoencoder variants used two graph-convolutional encoding stages with ReLU activation and dropout between successive transformations. A common hidden dimensionality of 64 and latent dimensionality of 64 were used in the reported multi-city experiments. GAE employed a GCN-based encoder, while ResGAE augmented the GCN representation with a residual projection from the input features to the latent space. GraphSAGE-AE replaced the GCN layers with GraphSAGE convolutional layers. VGAE used a shared GCN hidden layer followed by separate graph-convolutional transformations for the latent mean and log-variance. For all architectures, node attributes were reconstructed from the latent representation using a feed-forward feature decoder.

The deterministic autoencoders were optimized using a joint reconstruction objective consisting of binary cross-entropy for graph-structure reconstruction and mean-squared error for node-feature reconstruction, with equal weights assigned to the two components. For VGAE, the same reconstruction objective was augmented with a KL-divergence term weighted by $\beta_{\mathrm{KL}} = 0.1$. Adam optimization was used with a learning rate of 10^{-3}, weight decay of 10^{-5}, and dropout of 0.2. The main multi-city runs used a maximum of 50 training epochs with early stopping after 10 epochs without improvement in the validation objective. A validation ratio of 0.1 was used. Although GraphSAGE-AE employs GraphSAGE convolutional operators, the reported experiments used full-graph message passing rather than stochastic neighborhood sampling, ensuring that its training protocol remained directly comparable with the other autoencoder variants.

Before model training, the structural characteristics of the constructed multi-relational graphs were examined. Table 3 summarizes the graph statistics for each city, including the numbers of nodes, feature dimensions, union edges, and relation-specific edge counts. These statistics provide a quantitative description of the graph construction process and facilitate reproducibility of the experimental setting.

An explicit connectivity audit of the complete multi-relational graphs confirmed that no isolated nodes were present in any of the five study cities. Specifically, the number of isolated nodes was 0/23,656 in Baku, 0/38,391 in Turin, 0/64,361 in Vienna, 0/72,379 in Prague, and 0/91,503 in Kuala Lumpur. Thus, all 290,290 POIs participate in at least one relation in the final union graphs, and no isolated-node filtering, imputation, or special treatment was required prior to model training.

Table 3. Structural statistics of the constructed multi-relational urban POI graphs. The upper table reports general graph properties, while the lower table reports relation-specific edge counts. Geo_{75} and Geo_{250} denote the geo-close and geo-medium spatial relations constructed using 75 m and 250 m distance thresholds, respectively.

City	Nodes	Features	Union Edges	Avg. Degree
Baku	23,656	544	1,532,560	129.57
Kuala Lumpur	91,503	1629	7,344,925	160.54
Prague	72,379	1166	5,391,089	148.97
Turin	38,391	1202	3,027,408	157.71
Vienna	64,361	1297	4,851,622	150.76

City	Geo_{75}	Geo_{250}	Category	Type	Region
Baku	309,059	916,900	489,334	473,120	373,776
Kuala Lumpur	1,984,930	4,286,950	1,846,854	1,830,060	1,840,718
Prague	1,348,517	3,138,858	1,475,060	1,447,580	1,325,214
Turin	709,056	1,831,237	780,604	767,820	852,118
Vienna	1,134,529	2,860,942	1,317,492	1,287,220	1,288,996

After graph construction and representation learning, four complementary anomaly scoring perspectives were computed: reconstruction-based, cluster-based, neighborhood-consistency, and embedding-conflict scores. For each random seed, anomaly rankings were generated independently for every scoring perspective. The top-K anomalies were subsequently extracted and used to evaluate stability across seeds, construct consensus anomaly sets, compare agreement between scoring perspectives, and perform baseline and cross-city analyses.

Unless otherwise stated, consensus anomalies were derived from the multi-seed consensus framework described in Section 3.8. All experiments were performed using identical anomaly selection settings across cities to ensure comparability of the obtained results.

The main GAE, ResGAE, VGAE, and GraphSAGE-AE experiments were executed using GPU-enabled Google Colab runtimes. Because accelerator allocation in Google Colab is session dependent and the exact accelerator metadata from the original experimental sessions was not retained, the historical runs are not retrospectively attributed to a specific unverified GPU model. For reproducibility, the computational environment independently verified during the revision stage consisted of an NVIDIA Tesla T4 GPU with 14.56 GB of reported GPU memory, CUDA 12.8, PyTorch 2.11.0+cu128, and PyTorch Geometric 2.8.0.post1. Randomness was controlled using fixed experiment seeds for Python 3.13.15, NumPy 2.1.3, and PyTorch, including CUDA random-number generation where applicable.

As additional empirical context regarding computational cost, the GAD-NR baseline experiments introduced during revision required approximately 4916, 5026, 8163, 9204, and 9674 s for Baku, Turin, Vienna, Prague, and Kuala Lumpur, respectively. The corresponding adapted L-MAG-style experiments required approximately 61, 113, 211, 219, and 302 s. These measurements are provided to illustrate the computational scale of the additional baseline evaluation and should not be interpreted as direct runtime comparisons with the original graph-autoencoder experiments, because the baseline implementations and the original experiments were not executed under identical computational environments.

4.2. Multi-Seed Stability Results

Before constructing consensus anomaly sets, it is essential to verify that the detected anomalies remain stable under different random initializations. To evaluate this property, each graph autoencoder was trained using five independent random seeds, and the overlap between the corresponding Top-K anomaly sets was quantified using the intersection ratio

defined in Section 3.7. This analysis provides an initial assessment of whether the discovered anomalies represent persistent graph irregularities or merely artifacts of stochastic training dynamics.

Figure 3 illustrates a representative example of the stability analysis. The consistently high agreement values observed across all seed pairs indicate that the ranking of anomalous nodes remains largely preserved despite variations in random initialization. This behavior suggests that the learned graph representations capture stable structural characteristics of the urban environment rather than seed-specific optimization effects.

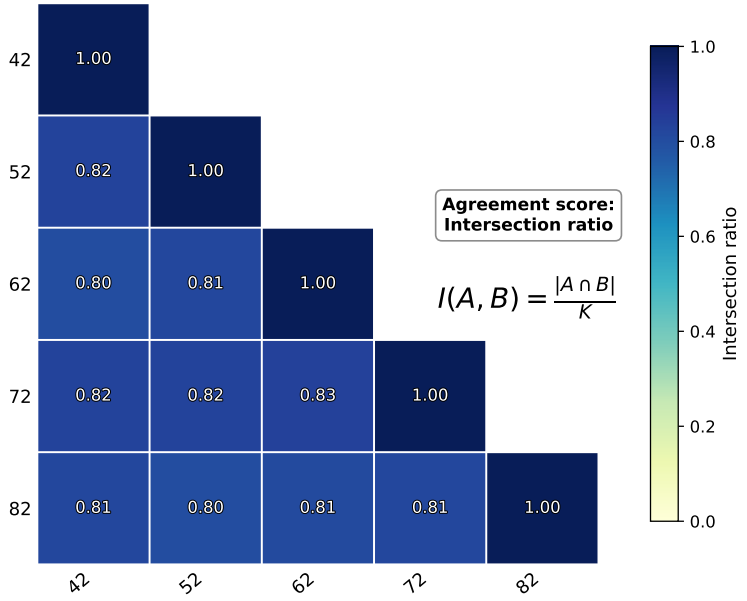


Figure 3. Representative seed-stability heatmap for the Vienna urban graph obtained using the ResGAE model and the neighborhood-consistency scoring perspective. Each cell reports the intersection ratio between the Top-K anomaly sets produced by two independent random seeds. High agreement across seed pairs indicates strong reproducibility of the detected anomalies.

The aggregated results reported in Table 4 further support the robustness of the anomaly discovery process across geographically distinct cities. Several model–score combinations consistently achieved stability values above 0.80, indicating a high degree of cross-seed consistency. In particular, the conflict-based and neighborhood-consistency perspectives exhibited the strongest stability, with ResGAE achieving average overlap ratios of 0.844 and 0.834, respectively. These findings indicate that anomalies identified by these perspectives are repeatedly recovered across independent training runs, providing stronger evidence that they correspond to consistent anomaly patterns. In contrast, cluster-based scoring generally produced lower stability values, suggesting a greater sensitivity to variations in the learned latent representations.

Table 4. Average seed-stability results across five cities. Values represent the mean pairwise intersection ratio between Top-K anomaly sets obtained from different random seeds.

Model	Scoring Perspective	Mean Stability
ResGAE	Conflict	0.844 ± 0.032
ResGAE	Neighbor	0.834 ± 0.039
GAE	Neighbor	0.815 ± 0.071
GAE	Conflict	0.812 ± 0.056
SAGEAE	Reconstruction	0.759 ± 0.202
VGAE	Reconstruction	0.755 ± 0.198

The stability analysis indicates that the anomaly rankings are not dominated by random initialization effects and remain largely consistent across multiple seeds. Having established substantial cross-seed stability in the detected anomaly rankings, the next step is to aggregate anomaly evidence across independent runs in order to identify the most reliable and persistent urban irregularities through the proposed consensus discovery framework.

4.3. Consensus Discovery Results

The stability analysis presented in the previous subsection showed that the detected anomalies remain largely consistent across different random seeds. Building upon this observation, a seed-level consensus framework was applied independently to each model–score configuration to aggregate anomaly evidence from repeated training runs and identify the most reliable urban irregularities. The resulting consensus sets provide a reproducible representation of anomalies within each configuration and form the basis for the analyses presented in this subsection.

Table 5 summarizes the average size of the consensus anomaly sets obtained for each city. Despite differences in urban structure, graph density, and POI composition, all cities produced substantial consensus sets, indicating that a large proportion of anomalous nodes were repeatedly identified across independent training runs. Baku and Turin exhibited the largest consensus sets, while Kuala Lumpur showed comparatively lower agreement, suggesting greater variability in the anomalies retained across different model–score configurations.

Table 5. Average within-configuration consensus anomaly set sizes across the five urban graphs. For each city, values are averaged over all model–score configurations, where consensus is computed across the five random seeds within each configuration.

City	Mean Consensus Size
Baku	188.9
Turin	185.4
Vienna	175.8
Prague	168.6
Kuala Lumpur	146.3

A more detailed analysis revealed notable differences between anomaly scoring perspectives. Reconstruction-based and conflict-based scores produced the largest and most stable consensus sets, whereas cluster-based scoring consistently generated smaller consensus sets. This observation is consistent with the stability results reported in Section 4.2, where cluster-based scores also exhibited the lowest seed agreement. Together, these findings suggest that reconstruction and conflict perspectives capture more persistent structural irregularities, while cluster-based anomalies are more sensitive to variations in the learned latent representations.

The within-configuration consensus framework successfully filters seed-specific detections and retains anomalies that repeatedly emerge across independent runs. As a result, the obtained consensus anomaly sets provide a more reliable basis for subsequent analyses. The next subsection investigates whether these consensus anomalies are also identified by established anomaly detection baselines, providing an additional perspective on their robustness and distinctiveness.

4.4. Comparison with Baseline Detectors

To further characterize the discovered consensus anomalies, we compare them with both established and more recent unsupervised anomaly detection baselines. The estab-

lished comparison includes LOF, DOMINANT, and COLA, representing density-based, graph-reconstruction, and contrastive anomaly-detection perspectives, respectively. In response to the need for comparison with more recent graph anomaly detection approaches, we additionally evaluate GAD-NR, a neighborhood-reconstruction graph anomaly detector [33], and an adapted lightweight implementation following the L-MAG design proposed in recent graph contrastive anomaly-detection research [40]. We refer to the latter as L-MAG-style throughout the experimental discussion because our implementation adapts the lightweight contrastive design to the present large-scale urban POI graphs rather than claiming an exact reproduction of the original implementation.

Because external node-level anomaly labels are unavailable for the five urban datasets, these comparisons should not be interpreted as conventional accuracy rankings among anomaly detectors. Instead, all methods are evaluated under a common Top-200 anomaly budget, and their overlap with the proposed consensus candidates is used to examine whether detectors based on different anomaly assumptions recover similar or complementary urban irregularity populations.

LOF and DOMINANT exhibited relatively low agreement with the discovered consensus anomalies. This behavior can be explained by the fundamentally different anomaly assumptions adopted by these methods. LOF detects local density deviations, while DOMINANT primarily identifies nodes with high reconstruction errors in graph structure and node attributes. In contrast, the proposed framework emphasizes anomalies that remain stable across multiple random initializations and persist under different anomaly-scoring perspectives. Consequently, the limited overlap suggests that the proposed framework emphasizes anomaly patterns that differ from those identified by density-based and reconstruction-based detectors, reflecting its stability- and consensus-oriented selection strategy.

Among the original three baselines, COLA exhibited the highest agreement with the discovered consensus anomalies. Unlike LOF and DOMINANT, COLA relies on contrastive representation learning, which emphasizes discrepancies between local graph context and learned node representations. This objective is conceptually closer to the neighborhood inconsistency and conflict-oriented perspectives employed by the proposed framework, which is consistent with its higher overlap relative to the other original baselines.

As shown in Table 6, agreement with the original baseline detectors varies substantially across the five study cities. COLA produces the largest average number of shared anomaly candidates and the highest mean coverage among these three methods, whereas LOF and DOMINANT show lower agreement. The relatively large cross-city standard deviations, particularly for COLA, indicate that baseline agreement is strongly dependent on urban context. Importantly, these standard deviations represent dispersion across cities and should not be interpreted as random-seed uncertainty.

Table 6. Agreement between the proposed consensus anomalies and representative baseline detectors across the five study cities. Values are reported as the mean ± standard deviation across cities ($n = 5$). Coverage denotes the proportion of consensus anomalies also identified by each baseline. The reported standard deviations therefore characterize cross-city variation in baseline agreement rather than repeated-run variability.

Baseline	Shared Nodes	Coverage (%)	Overlap F1 (%)
COLA	56.8 ± 37.9	28.7 ± 19.0	11.2 ± 7.1
LOF	31.2 ± 12.0	15.4 ± 5.3	12.8 ± 4.7
DOMINANT	24.5 ± 9.2	12.1 ± 4.4	10.3 ± 3.9

The comparison was subsequently extended with the two recent graph-based baselines. GAD-NR [33] reconstructs neighborhood information to identify nodes whose local graph

context is difficult to reproduce. Under the common Top-200 comparison budget, GAD-NR showed very limited overlap with the stability-based consensus candidates, with a five-city mean intersection of 0.4 nodes, mean consensus recovery of 0.20%, and mean Jaccard similarity of 0.0010. This result should not be interpreted as inferior anomaly-detection accuracy; rather, it indicates that the neighborhood-reconstruction objective identifies a substantially different anomaly population from the multi-model, multi-score consensus criterion used in the proposed framework.

As shown in Table 7, the adapted L-MAG-style contrastive baseline [40] exhibited greater overlap with the consensus candidates than GAD-NR. Across the five cities, the mean intersection was 27.6 nodes, corresponding to a mean consensus recovery of 14.00% and a mean Jaccard similarity of 0.0762. The degree of agreement remained city dependent: the largest recoveries were observed in Baku (23.86%) and Prague (21.21%), whereas Turin, Vienna, and Kuala Lumpur yielded 7.50%, 10.05%, and 7.37%, respectively. Thus, even a recent contrastive graph anomaly detector reproduces only part of the stable consensus population.

Table 7. Agreement between the proposed consensus candidates and the adapted L-MAG-style contrastive baseline under a common Top-200 anomaly budget. Recovery denotes the percentage of consensus candidates recovered by the baseline. The L-MAG-style implementation adapts the lightweight contrastive design to the present urban graphs and is not presented as an exact reproduction of the original implementation.

City	Shared Nodes	Recovery (%)	Jaccard
Baku	47	23.86	0.1343
Turin	15	7.50	0.0390
Vienna	20	10.05	0.0528
Prague	42	21.21	0.1180
Kuala Lumpur	14	7.37	0.0372
Mean	27.6	14.00	0.0762

Taken together, the baseline comparisons reveal substantial complementarity among anomaly definitions. Density deviation, graph reconstruction, neighborhood reconstruction, contrastive representation learning, and stability-aware consensus do not identify identical sets of urban POIs. The proposed framework is therefore not intended to replace these detectors on the basis of an accuracy claim that cannot be established without external anomaly labels. Its distinguishing objective is instead to identify candidates that persist across heterogeneous embedding architectures, anomaly-scoring perspectives, and stochastic executions. The partial overlap with both established and recent baselines indicates that some consensus candidates are supported by alternative anomaly criteria, while a substantial proportion remains specific to the stability-oriented selection mechanism.

Baseline overlap provides complementary evidence but does not constitute ground-truth validation of the detected anomalies. We therefore complement these comparisons with parameter-sensitivity and contextual validation analyses. The following subsection first evaluates whether the consensus selections remain robust under variations in candidate-set definition, agreement threshold, seed count, and spatial graph-construction parameters.

4.5. Sensitivity Analysis

The preceding experiments evaluate stability across repeated random initializations. We further examine whether the reported consensus results depend strongly on the main design choices used in candidate selection, consensus formation, and graph construction.

Specifically, the sensitivity analysis considers: (i) the fixed Top-K candidate-set size and consensus threshold T; (ii) city-size-normalized candidate sets defined as a proportion of the total number of POIs; (iii) the number of random seeds contributing to consensus formation; and (iv) the spatial thresholds used during graph construction. These analyses are intended to assess robustness rather than to identify uniquely optimal hyperparameters. The joint sensitivity to Top-K and the consensus threshold T is summarized in Table 8.

Table 8. Sensitivity analysis of the consensus mechanism under different Top-K and consensus threshold settings. Jaccard similarity measures the agreement between adjacent parameter configurations, while the last column summarizes the corresponding change in the average consensus set size across the five evaluated cities.

Analysis	Compared Settings	Jaccard Similarity	Consensus Size Trend
Top-K	$K = 100$ vs. $K = 150$ ($T = 3$)	0.696 ± 0.026	$720.8 \rightarrow 1037.0$
Top-K	$K = 150$ vs. $K = 200$ ($T = 3$)	0.769 ± 0.009	$1037.0 \rightarrow 1349.4$
Top-K	$K = 200$ vs. $K = 250$ ($T = 3$)	0.819 ± 0.011	$1349.4 \rightarrow 1649.2$
Top-K	$K = 250$ vs. $K = 300$ ($T = 3$)	0.844 ± 0.015	$1649.2 \rightarrow 1957.0$
Threshold	$T = 2$ vs. $T = 3$ ($K = 200$)	0.794 ± 0.076	$1727.0 \rightarrow 1349.4$
Threshold	$T = 3$ vs. $T = 4$ ($K = 200$)	0.844 ± 0.044	$1349.4 \rightarrow 1134.6$
Threshold	$T = 4$ vs. $T = 5$ ($K = 200$)	0.878 ± 0.028	$1134.6 \rightarrow 995.4$

Increasing Top-K gradually enlarges the consensus candidate pool, whereas increasing the consensus threshold produces smaller and more selective consensus sets. Despite the corresponding changes in set size, agreement between adjacent parameter configurations remains substantial. Mean Jaccard similarity ranges from 0.696 to 0.844 across adjacent Top-K settings and from 0.794 to 0.878 across adjacent threshold settings. Around the reference configuration, changing K from 150 to 200 yields a Jaccard similarity of 0.769 ± 0.009, while changing it from 200 to 250 yields 0.819 ± 0.011. Similarly, changing T from 2 to 3 and from 3 to 4 yields 0.794 ± 0.076 and 0.844 ± 0.044, respectively. The detected consensus sets therefore evolve progressively rather than changing abruptly around the reference setting.

Because a fixed $K = 200$ corresponds to different fractions of the POI population across cities, we additionally evaluated city-normalized candidate sets corresponding to 0.25%, 0.50%, 0.75%, and 1.00% of the nodes in each graph. The fixed $K = 200$, $T = 3$ configuration was used as the reference. Node-level consistency was evaluated through retention of the reference anomaly set, while Jensen–Shannon similarity was used separately to quantify similarity of the basic-category distributions.The corresponding results are reported in Table 9.

Table 9. Sensitivity to city-normalized anomaly candidate sets at $T = 3$. Reference retention denotes the proportion of anomalies from the fixed $K = 200$, $T = 3$ configuration that are retained, while category similarity denotes Jensen–Shannon similarity between the corresponding basic-category distributions.

Candidate Proportion	Reference Retention (%)	Category Similarity
0.25%	72.3	0.958
0.50%	92.1	0.975
0.75%	97.9	0.967
1.00%	100.0	0.958

The fixed $K = 200$ setting represents markedly different candidate proportions across the study cities, ranging from approximately 0.219% of the graph in Kuala Lumpur to 0.845% in Baku. Nevertheless, the 0.50% city-normalized setting retains approximately 92.1% of the reference anomalies at $T = 3$, while preserving a highly similar basic-category

composition with a mean Jensen–Shannon similarity of 0.975. Increasing the proportional candidate set to 0.75% raises reference retention to approximately 97.9%. At the 0.50% candidate setting, varying T from 2 to 5 retains approximately 80.7–94.4% of the reference anomalies, while category-composition similarity remains between 0.967 and 0.975. These results indicate that the qualitative composition of the consensus anomaly population is not specific to the use of a fixed Top-200 candidate set.

We also evaluated whether consensus formation depends critically on the use of five random seeds. Using the existing seed set $\{42, 52, 62, 72, 82\}$, all possible subsets were evaluated for $M = 2, 3, 4$, and 5 without retraining the models. Because the original $T = 3$ out of $M = 5$ rule corresponds to a 60% agreement requirement that cannot be represented exactly for every reduced seed count, the nearest practical integer thresholds were used: $T = 2$ for $M = 2$, $T = 2$ for $M = 3$, $T = 3$ for $M = 4$, and $T = 3$ for $M = 5$. The corresponding seed-count sensitivity results are summarized in Table 10.

Table 10. Sensitivity of the final cross-configuration consensus to the number of random-seed executions. Jaccard similarity and reference retention are computed relative to the full five-seed consensus.

Seeds M	Threshold T	Jaccard to $M = 5$	Reference Retention	Category Similarity
2	2	0.780	0.811	0.983
3	2	0.849	0.928	0.991
4	3	0.902	0.902	0.993
5	3	1.000	1.000	1.000

The seed-count analysis shows progressive convergence toward the five-seed reference. With two seeds, the mean Jaccard similarity is 0.780 and the mean reference retention is 0.811. Three seeds increase the mean Jaccard similarity to 0.849 and already recover approximately 92.8% of the five-seed reference anomalies, while category-composition similarity reaches 0.991. With four seeds, the mean Jaccard similarity rises further to 0.902 and category-composition similarity reaches 0.993. The same general convergence pattern is observed across all five cities; for $M = 4$, the city-level Jaccard similarities to the five-seed reference are approximately 0.936 for Baku, 0.919 for Turin, 0.894 for Prague, 0.892 for Vienna, and 0.871 for Kuala Lumpur. These results indicate that fewer seeds already recover most of the consensus structure, while five seeds provide an additional reproducibility margin rather than representing a uniquely optimal choice.

Finally, we evaluated the structural sensitivity of the spatial graph-construction thresholds. As clarified in Section 3.3, the graphs used in the reported experiments employ 75 m for the geo-close relation and 250 m for the geo-medium relation. The close-distance threshold was varied over 50, 75, and 100 m while fixing the medium threshold at 250 m, and the medium-distance threshold was varied over 200, 250, and 300 m while fixing the close threshold at 75 m. This analysis evaluates changes in graph structure only; no GNN retraining was performed for these threshold perturbations.

Across the five cities, varying the close threshold from 50 to 100 m changed the mean spatial degree from 80.65 to 88.64, while the mean isolated-node rate remained unchanged at 0.436% and mean largest-connected-component coverage remained unchanged at 92.43%. Varying the medium threshold from 200 to 300 m produced a larger but controlled structural effect: mean degree ranged from 75.63 to 92.41, mean isolated-node rates remained below 0.65%, and mean largest-connected-component coverage ranged from 88.94% to 95.52%. The default 75/250 m configuration yielded a mean degree of 85.39, a mean isolated-node rate of 0.436%, and mean largest-connected-component coverage of 92.43%. These results indicate that the adopted spatial configuration lies within a structurally stable neighborhood and is not dependent on a narrowly tuned threshold choice.

Taken together, these analyses support the use of $K = 200$, $T = 3$, five random seeds, and the 75/250 m spatial configuration as practical reference settings for the reported experiments. We do not interpret these values as uniquely optimal hyperparameters. Rather, the sensitivity results show that the principal consensus structure and category composition remain broadly consistent under city-normalized candidate sets, neighboring consensus thresholds, reduced seed counts, and local perturbations of the graph-construction radii. The reference configuration is therefore retained primarily to provide a common and reproducible evaluation protocol across the five cities and model configurations. Having established robustness to these principal design choices, we next examine the semantic and contextual characteristics of the detected consensus anomalies.

4.6. Urban Contextualization Results

To assess whether the stable consensus anomalies correspond to interpretable urban contexts rather than merely high model scores, we conducted a contextual case study centered on Prague. Prague was selected because its final configuration uses ResGAE with the neighbor-consistency score, making local spatial context directly relevant to the interpretation of the detected anomalies. The analysis combines spatial visualization, semantic composition, relation-specific structural profiles, and comparison with category-matched non-consensus controls.

Figure 4 provides a spatial overview of the Prague consensus set. Rather than relying on visual inspection alone, however, we evaluated the structural context of every Prague consensus anomaly against semantically comparable non-consensus POIs. For each consensus node, up to 20 non-consensus controls sharing the same basic category were selected. This matching strategy reduces the possibility that apparent structural differences are driven only by category rarity. The contextual comparison focuses on three explicitly spatial relations: geo-close, geo-medium, and same-region-cell. For each node, empirical city-wide percentile positions were computed for these relations and summarized through the number of low-tail relations, the number of extreme spatial relations, and a continuous spatial-tail score. Higher spatial-tail scores indicate relation profiles located further toward the tails of the corresponding city-wide distributions.

Uncertainty in the matched-control comparisons was quantified at the consensus-node level using paired resampling procedures. For each reported mean difference between a consensus anomaly and its category-matched control set, 95% confidence intervals were estimated from 5000 bootstrap resamples of the paired node-level differences. In addition, the paired differences were evaluated using 5000 paired sign-flip permutations under the null hypothesis of no systematic paired difference.

All 198 Prague consensus anomalies could be included in the matched-control analysis. Relative to their category-matched controls, the consensus anomalies exhibited, on average, 1040 additional low-tail spatial relations, with a 95% bootstrap confidence interval of [0.878, 1.202]. They also exhibited 0.866 additional extreme spatial relations on average, with a 95% confidence interval of [0.701, 1.026], and a mean spatial-tail-score increase of 0.203, with a 95% confidence interval of [0.175, 0.229]. The descriptive comparison with the full non-consensus population showed the same general pattern: the mean spatial-tail score was 0.786 for the consensus anomalies compared with 0.499 for non-consensus POIs, while the mean number of low-tail spatial relations was 1.707 and 0.274, respectively. These results provide population-level evidence that the stable neighbor-based Prague anomalies occupy atypical local spatial contexts relative to semantically matched controls.

To further investigate the nature of these anomalies, we analyzed their dominant semantic categories. The most frequent consensus anomaly categories identified in Prague are summarized in Table 11.

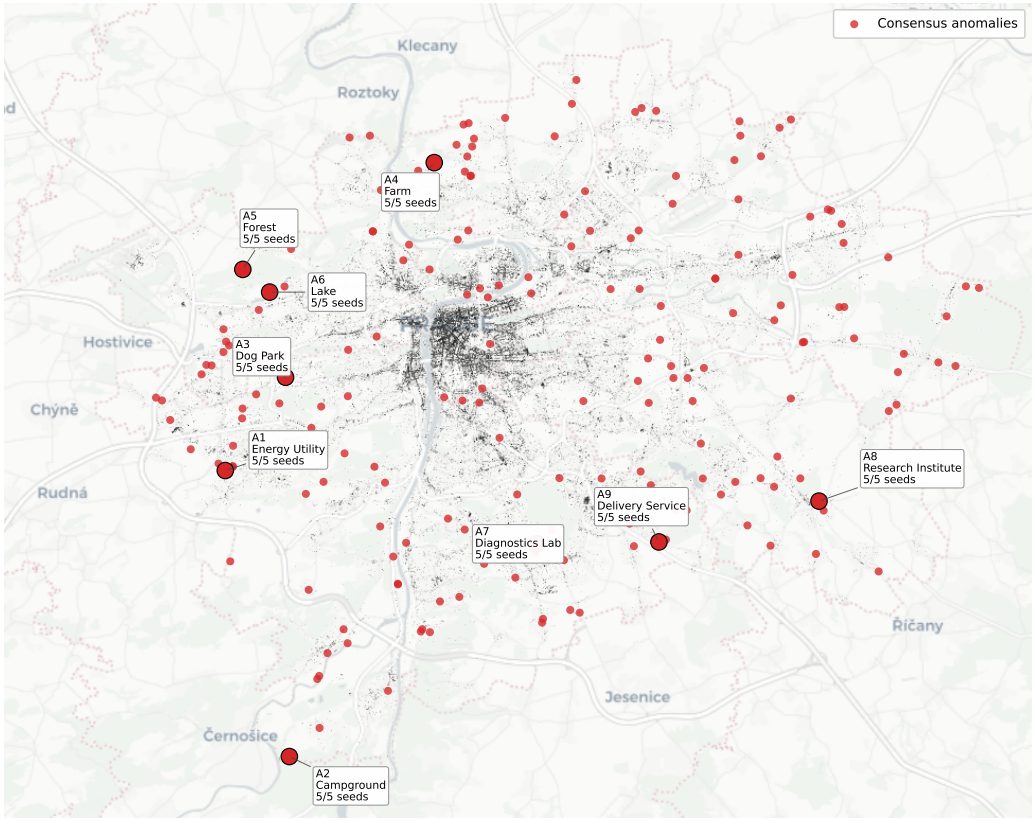


Figure 4. Spatial distribution of consensus anomalies identified in Prague. Red markers denote POIs repeatedly detected as anomalous across multiple runs. Annotated examples highlight representative anomalies from different urban regions.

Table 11. Most frequent semantic categories among the consensus anomalies identified in Prague.

Category	Frequency
Farm	17
Campground	12
Dog Park	9
Warehouse	9
Skate Park	7
Locksmith Service	7
Zoo	6
B2B Energy and Utility Service	6
Nature Reserve	6
Shipping or Delivery Service	5

The semantic composition provides an additional interpretive layer. Categories such as farms, campgrounds, nature reserves, zoos, and dog parks represent facilities whose local contexts may differ from those of more typical dense urban POIs, while warehouses, utility services, and shipping facilities reflect specialized infrastructure functions. However, the category-matched control design is important here: the contextual differences reported above cannot be explained solely by the fact that these categories are relatively uncommon, because each consensus anomaly is compared with non-consensus POIs from the same basic category.

To illustrate the relation profiles underlying the aggregate results, Table 12 reports three highly stable consensus anomalies that persisted across all five random-seed executions.

The Zoo and Resort examples lie close to the lower tails of all three spatial-relation distributions and exhibit spatial-tail scores above 0.97, while the Farm case also shows a strongly atypical multi-relation profile. These examples therefore connect the population-level matched-control result with concrete POIs whose stable anomaly status is accompanied by unusual local graph context.

Table 12. Representative stable Prague consensus anomalies and their empirical percentile positions in the three spatial relations used for contextual validation. Lower percentile values indicate relation degrees located closer to the lower tail of the corresponding city-wide distribution. The tail score summarizes the extremeness of the spatial-relation profile.

Category	Node	Geo-Close (%)	Geo-Medium (%)	Region-Cell (%)	Tail Score
Zoo	37796	1.99	1.35	0.97	0.971
Resort	72087	1.99	0.99	0.97	0.974
Farm	14756	6.11	3.05	5.58	0.903

As an additional robustness check, the same category-matched spatial-tail analysis was applied to the Baku consensus configuration. The same directional pattern was not observed: the mean differences between Baku consensus anomalies and their controls were -0.273 for low-tail spatial relations, -0.298 for extreme spatial relations, and -0.038 for the spatial-tail score. This contrast is informative rather than contradictory. Prague uses a neighborhood-consistency score that directly targets local neighborhood disagreement, whereas the selected Baku configuration uses a conflict-based anomaly score and therefore represents a different notion of irregularity. The contextual signature should consequently be interpreted as dependent on the anomaly-scoring perspective rather than as a universal structural property of all consensus anomalies.

Overall, the Prague case study provides contextual validation that is independent of overlap with external anomaly detectors: stable neighbor-based consensus anomalies exhibit measurable spatial-relation differences relative to semantically matched controls. At the same time, the contrasting Baku result demonstrates an important limitation and cautions against interpreting a single contextual signature as universally applicable across anomaly-scoring objectives. Because comprehensive external anomaly labels remain unavailable, these analyses should be viewed as contextual evidence supporting interpretability rather than as proof of absolute anomaly-detection accuracy. The following subsection therefore examines a separate question: whether consensus anomaly candidates exhibit non-random correspondence across different urban environments.

4.7. Cross-City Consistency Analysis

To examine whether the detected consensus anomalies exhibit correspondence across different urban environments, we first evaluated several similarity-weighting schemes and subsequently introduced explicit reference baselines to determine whether the observed matches exceed random expectation or can be explained by simpler semantic or structural similarity mechanisms. The analysis was conducted over all 20 directed source–target city pairs. For each pair, the 50 highest-vote source consensus anomalies were used as retrieval queries, and Consensus Hit@10, Hit@20, and Hit@50 were computed according to whether at least one target-city consensus anomaly appeared among the corresponding Top-N retrieved POIs. The results obtained under the original similarity-weighting schemes are summarized in Table 13.

Table 13. Cross-city anomaly consistency under different similarity weighting schemes. Hit@N represents the proportion of source-city consensus anomalies that retrieve at least one target-city consensus anomaly within the Top-*N* most similar matches. Best-performing values among the evaluated weighting schemes are shown in bold.

Scheme	Hit@10 (%)	Hit@20 (%)	Hit@50 (%)
Profile-Focused	5.2	8.4	14.1
Context-Focused	5.8	9.7	16.5
Balanced	11.4	17.8	28.6
Embedding-Focused	13.9	21.5	33.2
Semantic-Focused	**23.0**	**29.3**	**39.4**

Among the original weighting schemes, semantic-focused similarity produced the highest correspondence rates, reaching 23.0% at Hit@10 and 39.4% at Hit@50, whereas profile- and context-focused schemes produced substantially lower retrieval rates. This pattern motivated a more explicit baseline analysis to separate the contributions of semantic compatibility, graph-structural similarity, and random target-set density.

Four additional reference strategies were therefore evaluated: random retrieval, exact-category matching, semantic-only similarity, and structural-only similarity. Random retrieval repeatedly ranks the full target-city POI population without using source-query information. Exact-category matching prioritizes target POIs sharing the source POI's category, while semantic-only similarity uses semantic descriptors without structural information. Structural-only similarity excludes category and semantic information and relies only on local spatial and graph-structural characteristics. Randomized tie-breaking was used where necessary, uncertainty was quantified at the source-query level using bootstrap resampling, and empirical randomization tests were used to compare the combined similarity method against random retrieval. The corresponding reference-strategy results are summarized in Table 14.

Table 14. Mean cross-city consensus retrieval performance across all 20 directed source–target city pairs. Hit@N denotes the proportion of source consensus queries retrieving at least one target-city consensus anomaly within the Top-*N* ranked POIs.

Retrieval Strategy	Hit@10	Hit@20	Hit@50
Random Retrieval	0.0418	0.0815	0.1885
Structural-Only	0.0490	0.0930	0.2150
Combined Similarity	0.1140	0.1640	0.2880
Exact-Category	0.2668	0.3335	0.4142
Semantic-Only	0.2736	0.3469	0.4388

The combined similarity method achieved mean Consensus Hit@10, Hit@20, and Hit@50 values of 0.114, 0.164, and 0.288, respectively, compared with 0.0418, 0.0815, and 0.1885 under random retrieval. Thus, the corresponding retrieval rates were approximately 2.73, 2.01, and 1.53 times the random reference at the three cutoffs. Empirical randomization tests indicated that the combined method exceeded random retrieval at all three depths ($p < 0.001$). This provides evidence that the observed cross-city correspondence is not explained solely by random retrieval or by the size of the target consensus sets.

However, the additional baselines also substantially qualify the interpretation of the cross-city results. Semantic-only similarity achieved mean Hit@10/20/50 values of 0.274/0.347/0.439, while exact-category matching achieved 0.267/0.334/0.414, both exceeding the combined similarity method. In contrast, structural-only retrieval remained considerably closer to random retrieval, particularly at Hit@10 and Hit@20. These findings

indicate that semantic and category compatibility are major contributors to the observed cross-city matches, whereas non-semantic local structural similarity alone explains a substantially smaller proportion of the correspondence.

The results therefore support a more limited conclusion than broad cross-city transferability. A non-random subset of stable anomaly candidates exhibits correspondence across the five studied cities, but the stronger performance of semantic-only and exact-category retrieval shows that semantic overlap is an important driver of this correspondence. Moreover, the random Hit@N reference varies with the proportion of POIs belonging to the target-city consensus set, so raw cross-city Hit@N values should be interpreted in the context of target-city consensus density. The present analysis consequently provides evidence of non-random cross-city correspondence, but it does not establish that matched POIs represent the same underlying urban phenomenon or that the learned anomaly representations are universally transferable across cities.

Taken together, the stability, consensus, baseline-comparison, contextual-validation, and cross-city analyses provide complementary evidence regarding the reproducibility and interpretability of the proposed framework. These findings and their implications for urban analytics are discussed in the following section.

5. Discussion

5.1. Main Findings

The results demonstrate that explicitly accounting for stochastic variability can improve the reliability of unsupervised urban anomaly discovery in multi-relational POI graphs. Across the five study cities, repeated training under different random seeds revealed that individual anomaly rankings can vary, whereas recurrent candidates can be identified more reliably through stability analysis and consensus filtering. The contribution of the proposed framework therefore lies not in introducing a new graph autoencoder or a new voting rule in isolation, but in integrating multi-relational urban graph construction, heterogeneous embedding architectures, complementary anomaly-scoring perspectives, repeated-seed stability assessment, and consensus-based reliability filtering into a unified urban anomaly-discovery methodology. This stability-oriented perspective is particularly relevant for reproducible graph-based analysis, where conclusions derived from a single stochastic execution may otherwise depend on initialization-specific model behavior [12].

The consensus mechanism complements this stability analysis by prioritizing POIs that repeatedly emerge under multiple model, scoring, and seed configurations. Reconstruction error, cluster deviation, neighborhood inconsistency, and relation-conditioned conflict capture different manifestations of irregularity and therefore need not produce identical rankings. Their integration is consequently intended to reduce dependence on any single anomaly definition rather than to assume that one scoring perspective is universally optimal. The sensitivity experiments further support this design choice: moderate variations in the candidate-set size, consensus threshold, number of random seeds, and spatial graph-construction thresholds produced gradual rather than abrupt changes in the retained anomaly populations. This behavior is consistent with the general motivation of consensus-based anomaly analysis, in which recurrent evidence across heterogeneous views can reduce sensitivity to individual detector behavior [27].

The baseline experiments provide an additional perspective on this complementarity. The consensus candidates exhibited only partial overlap with LOF, DOMINANT, COLA, GAD-NR, and the adapted L-MAG-style contrastive baseline. In particular, the limited overlap with GAD-NR and the greater, but still partial, correspondence with the L-MAG-style baseline indicate that neighborhood reconstruction, contrastive representation learning, and stability-aware consensus emphasize different anomaly populations. These results

should not be interpreted as evidence that the proposed framework is more accurate than the individual baselines, because externally verified node-level anomaly labels are unavailable. Instead, they show that the framework provides a distinct stability-oriented criterion for selecting anomaly candidates that is complementary to established and recent detector objectives.

The contextual analysis provides evidence beyond detector overlap for interpreting the retained candidates. In Prague, all 198 consensus anomalies were compared with non-consensus POIs from the same basic categories. The consensus population exhibited substantially more atypical spatial-relation profiles, including a mean spatial-tail score of 0.786 compared with 0.499 for the broader non-consensus population and a matched-control mean difference of 0.203 with a 95% bootstrap confidence interval of [0.175, 0.229]. Representative stable cases such as a zoo, resort, and farm also occupied the lower tails of multiple spatial-relation distributions. Because controls were matched by category, these differences cannot be attributed solely to the rarity of the corresponding semantic classes. The contrasting Baku analysis, however, did not reproduce the same spatial-tail signature. This result emphasizes that contextual manifestations of irregularity depend on the anomaly-scoring perspective: Prague was selected using neighborhood inconsistency, whereas the Baku configuration emphasizes relation conflict. The contextual analysis should therefore be interpreted as supporting evidence for the urban interpretability of specific anomaly notions rather than as universal ground-truth validation.

The cross-city experiments further refine the interpretation of the framework. The combined retrieval strategy exceeded random retrieval across all evaluated depths, providing evidence that cross-city correspondence among consensus candidates is not entirely attributable to chance or target-set size. However, semantic-only and exact-category retrieval achieved higher correspondence than the combined strategy, whereas structural-only retrieval remained considerably closer to the random reference. These findings indicate that semantic compatibility is an important driver of the observed cross-city matches. Consequently, the results support non-random cross-city correspondence among a subset of anomaly candidates, but they do not establish broad transferability of anomaly patterns or universal cross-city generalization. Differences in POI coverage, category composition, urban structure, and consensus-set density remain relevant sources of variation and motivate a deliberately cautious interpretation of cross-city consistency.

From a smart-city perspective, this stability-oriented output is most naturally viewed as a screening and decision-support mechanism rather than an autonomous declaration that a facility is erroneous or improperly located. For example, municipal data-management teams could use repeatedly detected POIs to prioritize records for POI data-quality assessment, identifying facilities whose semantic attributes or relational context warrant verification. Urban planners could similarly use stable neighborhood- or conflict-based candidates as a first-stage screening tool for facility-layout assessment or land-use anomaly investigation. In such applications, the consensus vote, scoring perspective, semantic category, and relation-specific context provide complementary evidence that can guide human inspection. The framework therefore aims to reduce the search space for expert assessment while retaining human and domain-specific validation as the final decision stage.

5.2. Limitations

Several limitations should be considered when interpreting the results. First, the empirical evaluation is restricted to five cities. Although the selected cities differ substantially in geographic context, graph size, feature dimensionality, and urban composition, they cannot represent the full diversity of urban forms, mapping practices, and development patterns observed worldwide. The multi-city experiments therefore provide evidence

across heterogeneous urban environments rather than proof of universal generalizability. Evaluation on a substantially larger and more systematically sampled collection of cities would be required to establish broader geographic robustness.

Second, externally verified node-level labels identifying genuine urban anomalies are unavailable. Consequently, conventional detection metrics such as precision, recall, or AUROC against a validated ground truth cannot be reported, and neither consensus frequency nor agreement with another detector should be interpreted as proof that a POI constitutes a true urban anomaly. We addressed this limitation through complementary forms of evidence, including multi-seed stability analysis, consensus sensitivity analysis, comparisons with established and recent anomaly detectors, category-matched contextual validation, and cross-city retrieval baselines. These analyses strengthen the assessment of reproducibility and interpretability but do not replace externally validated anomaly labels.

Third, the analysis depends on the quality, completeness, and semantic consistency of Overture Maps POI data. Differences among cities in POI coverage, category granularity, attribute completeness, duplication, and regional mapping practices may alter both graph construction and anomaly characterization. Moreover, categorical encoding produces different feature dimensionalities across cities, reflecting differences in the observed urban data vocabularies. Some detected irregularities may therefore reflect characteristics of the underlying digital map representation in addition to characteristics of the physical urban environment. This distinction is particularly important when the framework is used for POI data-quality assessment, where a detected anomaly may indicate either a genuine unusual facility context or a record that warrants verification.

Fourth, graph construction necessarily depends on modeling choices. The spatial relations use fixed distance thresholds of 75 m and 250 m, while regional relations depend on the adopted 200 m $\times$ 200 m grid partition. Although the spatial-threshold sensitivity analysis indicates that moderate perturbations around the selected spatial radii preserve the overall structural conclusions, these settings may not be equally appropriate for every urban morphology or POI density. Dense city centers, suburban areas, and sparsely mapped regions may operate at different characteristic spatial scales. Future extensions could therefore consider density-adaptive or data-driven neighborhood definitions rather than globally fixed spatial thresholds.

Fifth, the cross-city analysis should not be interpreted as demonstrating broad anomaly transferability. Although the combined retrieval strategy significantly exceeded random retrieval, semantic-only and exact-category baselines achieved stronger correspondence, showing that shared semantic categories are an important driver of cross-city matches. In addition, random Hit@N depends partly on the relative size of the target-city consensus population, while differences in POI coverage and category composition may further influence retrieval rates. Cross-city matches therefore demonstrate non-random correspondence among subsets of anomaly candidates, but they do not establish that matched POIs represent identical underlying urban phenomena or that anomaly representations learned in one city will generalize without qualification to another.

Sixth, the contextual validation provides evidence for interpretability but remains dependent on the anomaly perspective being examined. In Prague, neighbor-consistency consensus anomalies exhibited substantially more atypical spatial-relation profiles than category-matched controls, whereas the same directional spatial-tail signature was not reproduced for the conflict-based Baku configuration. This contrast indicates that different scoring perspectives capture different forms of irregularity. Consequently, no single contextual criterion should be treated as a universal external validator for all anomaly types, and broader validation would benefit from multiple domain-specific criteria and, where feasible, expert assessment.

Finally, the framework introduces additional computational cost relative to a single-model, single-seed anomaly detector because robustness is assessed across multiple graph autoencoder architectures, anomaly-scoring perspectives, and random initializations. The city graphs contain tens of thousands of POIs and, depending on the city, large numbers of relational edges, making repeated graph representation learning the principal scalability consideration. Although individual model–seed runs are independent and can therefore be parallelized, the current study does not provide a comprehensive runtime or memory-complexity benchmark across alternative hardware configurations. Scaling the framework to substantially larger metropolitan graphs or much larger city collections may require relation-aware edge sampling, mini-batch training, distributed execution, or more selective model–score ensembles.

5.3. Future Work

Several research directions emerge from the present findings and limitations. First, future work should extend the evaluation to a substantially larger and more systematically selected collection of cities spanning additional continents, urban morphologies, and levels of POI coverage. Such an extension would enable a more rigorous assessment of the geographic robustness of the discovered anomaly patterns and help distinguish genuinely recurring cross-city characteristics from correspondences driven primarily by shared semantic categories or dataset composition.

A second important direction concerns external validation. Because verified node-level urban anomaly labels are unavailable in the present study, future evaluations could incorporate expert assessment by urban planners or geospatial-data specialists, targeted manual verification of high-consensus candidates, or application-specific reference information. Such validation would help determine whether stable anomaly candidates correspond to mapping inconsistencies, unusual facility layouts, atypical land-use contexts, or other practically relevant forms of urban irregularity.

Third, the current framework analyzes static urban snapshots. Incorporating temporally resolved POI, mobility, or land-use information would enable the investigation of emerging, evolving, and disappearing irregularities and would provide a richer characterization of urban dynamics beyond the static setting considered here.

Fourth, future investigations could explore dynamic and explicitly relation-aware graph architectures capable of modeling temporal, semantic, spatial, and regional interactions without collapsing all relation types into a unified topology during representation learning. This direction could provide richer relation-specific representations and allow the contribution of individual urban relations to anomaly formation to be studied more directly. Adaptive or density-aware spatial neighborhood definitions could also be investigated as alternatives to globally fixed distance thresholds.

Fifth, computational scalability should be investigated systematically as the framework is extended to larger metropolitan graphs and larger collections of cities. Because the present stability-oriented design requires repeated training across multiple architectures and random initializations, future implementations could investigate mini-batch graph training, neighborhood or relation-aware edge sampling, graph partitioning, sparse computation, and parallel execution to reduce computational requirements while preserving anomaly-ranking stability.

Finally, integrating additional urban data modalities, such as mobility flows, traffic measurements, remote sensing imagery, or socio-economic indicators, may improve anomaly interpretability and enable a more holistic characterization of urban irregularities. Combining these complementary information sources with the existing POI-based representation could also help distinguish anomalies arising from incomplete or inconsistent map data from those reflecting unusual characteristics of the underlying urban environment.

6. Conclusions

This paper presented a stability-aware consensus framework for unsupervised urban anomaly discovery using multi-relational POI graphs constructed from Overture Maps data. Unlike conventional anomaly detection approaches that rely on a single detector output, the proposed framework integrates multiple graph autoencoder architectures, multiple anomaly scoring perspectives, and repeated random initializations to identify anomalies that remain consistently detectable across different analytical conditions. The resulting consensus mechanism prioritizes anomaly candidates that recur across heterogeneous analytical configurations, thereby providing a stability-oriented alternative to relying on the output of a single stochastic model execution.

Experiments conducted across five geographically distinct cities provided complementary evidence regarding the stability and interpretability of the proposed framework. Multi-seed evaluation and consensus-parameter sensitivity analyses showed that the retained anomaly populations remain reasonably consistent under repeated initialization and moderate variations in the consensus configuration. Comparisons with LOF, DOMINANT, COLA, GAD-NR, and an adapted L-MAG-style contrastive baseline revealed partial rather than complete overlap, indicating that stability-aware consensus selects anomaly populations that are complementary to those emphasized by individual detector objectives. Contextual validation further showed that the Prague neighbor-consistency consensus anomalies exhibited substantially more atypical spatial-relation profiles than category-matched non-consensus controls, while the contrasting Baku analysis demonstrated that such contextual signatures depend on the anomaly perspective being examined.

The cross-city experiments also showed that correspondence among consensus candidates is not explained entirely by random retrieval. However, semantic-only and exact-category retrieval produced stronger correspondence than the combined similarity strategy, indicating that shared semantic information contributes substantially to the observed cross-city matches. The findings therefore support non-random cross-city correspondence among subsets of anomaly candidates rather than broad transferability or universal generalization of anomaly patterns across cities.

Overall, the results support stability-aware consensus as a practical mechanism for identifying recurrent anomaly candidates in large urban POI graphs while making the uncertainty associated with stochastic model behavior explicit. By combining multi-relational graph construction, heterogeneous graph autoencoder architectures, complementary anomaly-scoring perspectives, repeated-seed stability assessment, consensus filtering, contextual validation, and controlled cross-city analysis, the framework provides a unified methodology for reproducibility-oriented urban anomaly discovery. In practical smart-city settings, the resulting candidates can be used as a screening layer for tasks such as POI data-quality assessment, facility-layout assessment, and land-use anomaly investigation, while domain expertise or external evidence remains necessary before interpreting a detected candidate as a verified urban irregularity. Future work will extend the evaluation to broader collections of cities and investigate expert annotation, external urban-planning and land-use information, adaptive graph construction, additional urban data modalities, and scalable training strategies for stronger external validation and geographic robustness.

Author Contributions: Conceptualization, E.V. and S.M.; methodology, E.V.; software, E.V.; validation, E.V. and S.M.; formal analysis, E.V.; investigation, E.V.; data curation, E.V.; writing—original draft preparation, E.V.; writing—review and editing, S.M. and D.A.; visualization, E.V.; supervision, D.A.; project administration, D.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received institutional support from ADA University and Politecnico di Torino. No specific grant number was associated with this research.

Data Availability Statement: The urban POI dataset used in this study was constructed from publicly available geospatial data provided by the Overture Maps Foundation. The processed dataset, graph construction pipeline, and embedding evaluation code can be made available by the corresponding author upon reasonable request to support research reproducibility.

Acknowledgments: The authors used OpenAI's ChatGPT (GPT-5) exclusively to assist in generating the conceptual layouts of Figures 1 and 2 based on author-developed methodology and experimental results. All scientific content, experimental design, analyses, interpretations, and conclusions were conceived, verified, and approved by the authors.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Chen, Y.; Li, X.; Cong, G.; Long, C.; Bao, Z.; Liu, S.; Gu, W.; Zhang, F. Points-of-Interest Relationship Inference with Spatial-enriched Graph Neural Networks. *arXiv* **2022**, arXiv:2202.13686. [CrossRef]
2. Ning, Y.; Liu, H.; Wang, H.; Zeng, Z.; Xiong, H. UUKG: Unified urban knowledge graph dataset for urban spatiotemporal prediction. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 62442–62456. [CrossRef]
3. Jin, X.; Tao, J.; Fu, K. Location Representation Learning for Urban Management with Heterogeneous Graph. In *Proceedings of the 2025 26th IEEE International Conference on Mobile Data Management (MDM), Irvine, CA, USA, 2–5 June 2025*; IEEE: New York, NY, USA, 2025; pp. 113–120.
4. Kim, N.; Yoon, Y. Effective Urban Region Representation Learning Using Heterogeneous Urban Graph Attention Network (HUGAT). *IEEE Access* **2025**, *13*, 102602–102612. [CrossRef]
5. Neira, M.; Murcio, R. Graph representation learning on street networks. *Isprs Int. J.-Geo-Inf.* **2025**, *14*, 284. [CrossRef]
6. Duan, P.; Ren, X.; Liu, Y. Multi-relational dynamic graph representation learning. *Neurocomputing* **2023**, *558*, 126688. [CrossRef]
7. Fang, Y.; Li, X.; Ye, R.; Tan, X.; Zhao, P.; Wang, M. Relation-aware graph convolutional networks for multi-relational network alignment. *Acm Trans. Intell. Syst. Technol.* **2023**, *14*, 37. [CrossRef]
8. Ekle, O.A.; Eberle, W. Anomaly detection in dynamic graphs: A comprehensive survey. *Acm Trans. Knowl. Discov. Data* **2024**, *18*, 192. [CrossRef]
9. Jin, M.; Koh, H.Y.; Wen, Q.; Zambon, D.; Alippi, C.; Webb, G.I.; King, I.; Pan, S. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10466–10485. [CrossRef]
10. Kamiński, B.; Prałat, P.; Théberge, F. An unsupervised framework for comparing graph embeddings. *J. Complex Netw.* **2020**, *8*, cnz043. [CrossRef]
11. Alahmari, S.S.; Goldgof, D.B.; Mouton, P.R.; Hall, L.O. Challenges for the repeatability of deep learning models. *IEEE Access* **2020**, *8*, 211860–211868. [CrossRef]
12. Desai, A.; Abdelhamid, M.; Padalkar, N.R. What is reproducibility in artificial intelligence and machine learning research? *AI Mag.* **2025**, *46*, e70004. [CrossRef]
13. Zhang, Q.; Gao, X.; Wang, H.; Huang, D.; Yiu, S.M.; Yin, H. HGAurban: Heterogeneous Graph Autoencoding for Urban Spatial-Temporal Learning. In Proceedings of the 34th ACM International Conference on Information and Knowledge Management, Seoul, Republic of Korea, 10–14 November 2025; pp. 4139–4148.
14. Li, S.; Zhou, J.; Liu, J.; Xu, T.; Chen, E.; Xiong, H. Multi-temporal relationship inference in urban areas. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; pp. 1316–1327.
15. Fan, C.; Yang, Y.; Mostafavi, A. Neural embeddings of urban big data reveal spatial structures in cities. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 409. [CrossRef]
16. Wang, H.; Che, X.; Chang, E.; Qu, C.; Zhang, G.; Zhou, Z.; Wei, Z.; Lyu, G.; Li, P. Similarity based city data transfer framework in urban digitization. *Sci. Rep.* **2025**, *15*, 10776. [CrossRef]
17. Zhang, H.; Zhou, Y.; Xu, H.; Shi, J.; Lin, X.; Gao, Y. Graph neural network approach with spatial structure to anomaly detection of network data. *J. Big Data* **2025**, *12*, 105. [CrossRef]
18. Du, X.; Yu, J.; Chu, Z.; Jin, L.; Chen, J. Graph autoencoder-based unsupervised outlier detection. *Inf. Sci.* **2022**, *608*, 532–550. [CrossRef]
19. Hu, Y.; Qu, A.; Work, D. Detecting extreme traffic events via a context augmented graph autoencoder. *ACM Trans. Intell. Syst. Technol. (TIST)* **2022**, *13*, 101. [CrossRef]
20. Tang, M.; Yang, C.; Li, P. Graph auto-encoder via neighborhood wasserstein reconstruction. *arXiv* **2022**, arXiv:2202.09025.

21. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 93–104.
22. Ding, K.; Li, J.; Bhanushali, R.; Liu, H. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining, Calgary, AB, Canada, 2–4 May 2019*; SIAM: Philadelphia, PA, USA, 2019; pp. 594–602.
23. Liu, Y.; Li, Z.; Pan, S.; Gong, C.; Zhou, C.; Karypis, G. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 2378–2392. [CrossRef]
24. Sun, Z.; Su, J.; Jeon, D.; Velasquez, A.; Song, H.; Niu, S. Reinforced contrastive graph neural networks (RCGNN) for anomaly detection. In *Proceedings of the 2022 IEEE International Performance, Computing, and Communications Conference (IPCCC), Austin, TX, USA, 11–13 November 2022*; IEEE: New York, NY, USA, 2022; pp. 65–72.
25. Carbone, F.; Cenedese, A.; Pizzi, C. Consensus-based anomaly detection for efficient heating management. In *Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), San Francisco, CA, USA, 4–8 August 2017*; IEEE: New York, NY, USA, 2017; pp. 1–7.
26. Li, L.; Qu, H.; Li, Z.; Zheng, J.; Tang, X.; Liu, P. Data reconstruction via consensus graph learning for effective anomaly detection in industrial iot. *IEEE Trans. Ind. Inform.* **2023**, *20*, 3996–4006. [CrossRef]
27. Yahaya, S.W.; Lotfi, A.; Mahmud, M. A consensus novelty detection ensemble approach for anomaly detection in activities of daily living. *Appl. Soft Comput.* **2019**, *83*, 105613. [CrossRef]
28. Shu, X.; Yang, H.; Wu, F.; Zhou, Y.; Xu, P.; Liu, X.; Guo, J.; Cai, Y.; Fan, J. An integrated spatiotemporal trust and consensus fusion framework for dam safety assessment with multi-sensor anomaly detection. *Comput.-Aided Civ. Infrastruct. Eng.* **2025**, *40*, 5625–5648. [CrossRef]
29. Jiang, M.; Hou, C.; Zheng, A.; Hu, X.; Han, S.; Huang, H.; He, X.; Yu, P.S.; Zhao, Y. Weakly supervised anomaly detection: A survey. *arXiv* **2023**, arXiv:2302.04549.
30. Zhang, K.; Lu, G.; Li, Y.; Xu, C. A Graph Autoencoder-based Anomaly Detection Method for Attributed Networks. In Proceedings of the 2023 5th International Conference on Natural Language Processing (ICNLP), Guangzhou, China, 24–26 March 2023; pp. 330–337. [CrossRef]
31. Sadrolhefazi, Z.S.; Zare, H.; Asghari, S.A.; Moradi, P.; Jalili, M. Unsupervised Graph-Based Anomaly Detection Using Variational Autoencoder. In *Proceedings of the 2024 IEEE International Conference on Data Mining Workshops (ICDMW), Abu Dhabi, United Arab Emirates, 9 December 2024*; IEEE: New York, NY, USA, 2024; pp. 734–741. [CrossRef]
32. Hamilton, W.; Ying, Z.; Leskovec, J. Inductive representation learning on large graphs. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1025–1035.
33. Roy, A.; Shu, J.; Li, J.; Yang, C.; Elshocht, O.; Smeets, J.; Li, P. GAD-NR: Graph Anomaly Detection via Neighborhood Reconstruction. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining, Mérida, Mexico, 4–8 March 2024; pp. 576–585. [CrossRef]
34. Zheng, L.; Birge, J.; Wu, H.; Zhang, Y.; He, J. Cluster aware graph anomaly detection. In Proceedings of the ACM on Web Conference 2025, Sydney, Australia, 28 April–2 May 2025; pp. 1771–1782.
35. Kloberdanz, E.; Kloberdanz, K.G.; Le, W. DeepStability: A study of unstable numerical methods and their solutions in deep learning. In Proceedings of the 44th International Conference on Software Engineering, Pittsburgh, PA, USA, 21–29 May 2022; pp. 586–597.
36. Nawaz, A.; Khan, S.S.; Ahmad, A. Ensemble of autoencoders for anomaly detection in biomedical data: A narrative review. *IEEE Access* **2024**, *12*, 17273–17289. [CrossRef]
37. Li, Y.; Huang, W.; Cong, G.; Wang, H.; Wang, Z. Urban region representation learning with openstreetmap building footprints. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; pp. 1363–1373.
38. Xu, R.; Chen, Z.; Li, F.; Zhou, C. Identification of urban functional zones based on POI density and marginalized graph autoencoder. *Isprs Int. J.-Geo-Inf.* **2023**, *12*, 343. [CrossRef]
39. Khoshraftar, S.; An, A. A survey on graph representation learning methods. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 19. [CrossRef]
40. Liu, Z.; Cao, C.; Tao, F.; Sun, J. Revisiting graph contrastive learning for anomaly detection. In *Proceedings of the ECAI 2023: 26th European Conference on Artificial Intelligence–Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023), Kraków, Poland, 30 September–4 October 2023*; SAGE Publications: London, UK, 2023; pp. 1585–1592.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.