DOI: https://doi.org/10.36910/6775-2524-0560-2026-62-12
UDC 004.275
Arzubov Maksym, PhD student, Assistant Lecturer
https://orcid.org/0000-0002-4592-0965
Narushynska Olga, PhD student, Assistant Lecturer
https://orcid.org/0009-0002-2994-6556
Lviv Polytechnic National University, Lviv, Ukraine

RETHINKING CLUSTER QUALITY EVALUATION FOR LARGE-SCALE GEOSPATIAL DATA

Arzubov M., Narushynska O. Rethinking Cluster Quality Evaluation for Large-Scale Geospatial Data. This paper advances cluster quality evaluation for large-scale geospatial data by challenging the adequacy of classical validation metrics based on global or pairwise distance computations. Although widely used, such metrics fail to scale computationally and inadequately capture spatial heterogeneity, local density variation, and multi-scale spatial organization. We argue that cluster quality evaluation in this context requires a shift from global distance-based reasoning toward locality-driven assessment. To this end, we propose a scalable evaluation framework grounded in local spatial structure, which quantifies neighborhood-level compactness, density consistency, and spatial coherence. Local quality indicators are computed within spatial neighborhoods and aggregated into a global assessment without exhaustive pairwise analysis. The proposed framework is algorithm-agnostic and applicable to crisp, fuzzy, and hierarchical clustering results. Experimental validation on large real-world geospatial datasets demonstrates that the proposed approach delivers stable, interpretable, and scale-consistent quality estimates while significantly reducing computational complexity compared to conventional internal validation measures. Furthermore, the results reveal improved robustness to non-uniform spatial distributions and density gradients, exposing limitations of traditional metrics at global scales. Overall, the proposed framework constitutes a more appropriate evaluation paradigm for clustering large-scale geospatial data in modern analytical and interactive mapping environments.
Keywords: cluster quality evaluation, geospatial clustering, large-scale geospatial data, spatial heterogeneity, local spatial structure, scalability, clustering validation, density-aware evaluation, pairwise distance, clustering.

Арзубов М.В., Нарушинська О.О. Альтернативний підхід до оцінювання якості кластеризації великих геопросторових даних. У статті запропоновано новий підхід до оцінювання якості кластеризації великих геопросторових даних, спрямований на подолання обмежень традиційних метрик валідації, що ґрунтуються на глобальних або попарних відстанях. Незважаючи на широке застосування, такі метрики є обчислювально неефективними та недостатньо чутливими до просторової неоднорідності, локальних змін густини та багатомасштабної організації геоданих. Обґрунтовується необхідність переходу від глобального дистанційного аналізу до локально орієнтованого оцінювання, яке краще відображає просторову структуру даних. Запропонований масштабований фреймворк оцінювання базується на використанні локальної просторової структури та передбачає обчислення показників компактності, узгодженості густини та просторової зв'язаності в межах локальних околів із подальшою агрегацією результатів у глобальну оцінку без виконання повного попарного аналізу. Розроблений підхід є незалежним від конкретного алгоритму кластеризації та може застосовуватися до чітких, нечітких ієрархічних кластерних моделей. Експериментальні дослідження на великих наборах геопросторових даних підтверджують стабільність, інтерпретованість і масштабну узгодженість отриманих оцінок при істотному зменшенні обчислювальних витрат. Отримані результати свідчать про підвищену стійкість запропонованого підходу до нерівномірних просторових розподілів і градієнтів густини, що дозволяє розглядати його як більш адекватну парадигму оцінювання кластеризації великих геоданих у сучасних аналітичних і картографічних середовищах.
Ключові слова: оцінювання якості кластеризації, геопросторова кластеризація, великі геопросторові дані, просторова неоднорідність, локальна просторова структура, масштабованість, валідація кластеризації, оцінювання з урахуванням густини, попарні відстані, кластеризація.

Statement of a scientific problem. Clustering plays a central role in the analysis of geospatial data by enabling the identification of spatial patterns, structural regularities, and latent spatial groupings. In recent years, the scale and complexity of geospatial datasets have increased significantly due to advances in sensing technologies, location-aware services, and large-scale spatial data collection. As a result, clustering outcomes are increasingly used not only for offline analysis, but also as a basis for downstream analytical tasks, decision support, and interactive spatial exploration.

Despite the extensive development of clustering algorithms, the problem of evaluating clustering quality in large-scale geospatial settings remains insufficiently addressed. Existing evaluation practices are largely inherited from general-purpose data analysis and are often applied to geospatial data without explicit consideration of spatial heterogeneity, scale dependency, and locality effects. In large and unevenly distributed spatial datasets, such practices may produce quality estimates that are difficult to interpret, unstable across scales, or impractical to compute under real-world constraints.

This gap highlights a fundamental scientific problem: how to formulate cluster quality evaluation in a manner that is compatible with the spatial nature of geospatial data and remains applicable under large-scale conditions. Addressing this problem requires an evaluation perspective that accounts for local spatial

organization and density variation, while avoiding assumptions that become invalid as data volume and spatial complexity increase.

Objectives of the Study:

- To formalize the limitations of classical cluster validation metrics when applied to large-scale geospatial data, with particular attention to scalability constraints, spatial heterogeneity, and sensitivity to local density variations.
- To develop a locality-driven cluster quality evaluation method that exploits local spatial structure to capture neighborhood-level compactness, density consistency, and spatial coherence without relying on exhaustive pairwise distance computations.
- To ensure algorithmic generality of the proposed evaluation approach, enabling its application to crisp, fuzzy, and hierarchical clustering results independently of the underlying clustering algorithm.
- To experimentally validate the proposed method on large real-world geospatial datasets, assessing computational efficiency, robustness to non-uniform spatial distributions, and consistency of quality estimates across spatial scales.

Research analysis. Recent studies on cluster quality evaluation for large-scale geospatial data confirm the ongoing relevance of developing scalable and spatially meaningful validation methods. Classical internal cluster validity indices, including distance- and dispersion-based measures, remain widely applied due to their simplicity and algorithm-agnostic nature[1–5] . However, these metrics were designed for general-purpose data analysis and rely on global aggregation or pairwise distance computations, which limits their interpretability and computational feasibility for large, spatially heterogeneous geospatial datasets.

Density-based and hierarchical clustering methods, such as DBSCAN, OPTICS, and HDBSCAN, are extensively used in geospatial analysis to identify non-convex structures and variable-density clusters [6]–[10]. Despite their suitability for spatial data, cluster quality assessment in practice often still relies on general-purpose internal indices, which may inadequately reflect density variation, noise, and multi-scale spatial organization. Several recent works propose density-aware validation measures to address these issues [11–14], yet such approaches are frequently computationally demanding or tightly coupled to specific clustering paradigms.

Parallel research in GIScience emphasizes locality through spatial autocorrelation statistics and local indicators of spatial association [15–18]. These methods provide valuable insights into neighborhood-level spatial patterns but are not intended as algorithm-independent measures of clustering quality. Additionally, spatial aggregation effects and scale sensitivity, highlighted by the modifiable areal unit problem, further complicate reliable evaluation across different spatial resolutions [19].

Recent efforts to improve scalability focus on approximating or distributing classical validation metrics using sampling or parallel computing frameworks [20–30]. While these approaches significantly reduce computational costs, they largely preserve global distance-based evaluation logic and do not explicitly incorporate local spatial structure into the definition of cluster quality.

Overall, existing research demonstrates substantial progress in clustering algorithms and scalable computation but reveals a persistent gap in cluster quality evaluation methods that are simultaneously scalable, locality-aware, and applicable across different clustering paradigms. This gap motivates the development of evaluation frameworks based on local spatial structure and efficient spatial indexing, enabling robust and interpretable assessment of clustering quality in large-scale geospatial data.

The purpose of the work. The aim of the study is to develop a scalable and methodologically grounded approach to cluster quality evaluation for large-scale geospatial data. The work aims to address the limitations of conventional validation metrics that rely on global or pairwise distance computations and exhibit insufficient sensitivity to spatial heterogeneity and local density variation. In particular, the objective is to formulate an evaluation method that leverages local spatial structure to provide robust, interpretable, and scale-consistent quality estimates for clustering results. The proposed approach is intended to remain computationally feasible for large datasets and to be applicable across different clustering paradigms, including crisp, fuzzy, and hierarchical models, thereby supporting reliable geospatial analysis in modern large-scale analytical and mapping environments.

Presentation of the main material and substantiation of the obtained research results.

Large-scale geospatial data are inherently characterized by spatial heterogeneity, non-uniform density distributions, and scale-dependent spatial relationships. Under such conditions, global distance-based representations are insufficient to adequately describe the spatial organization of data points. To

address this limitation, cluster quality evaluation must be grounded in a formal representation of local spatial structure.

Let $X = \{x_1, x_2, \dots, x_n\}, x_i \in \mathbb{R}^2$ be a large-scale geospatial dataset, where each point $x_i$is defined by spatial coordinates in a metric space equipped with a distance function $d(\cdot,\cdot)$.

For each point x_i, a local spatial neighborhood $\mathcal{N}(x_i) \subset X$is defined as:

$$\mathcal{N}(x_i) = \{ x_j \in X \mid d(x_i, x_j) \leq r \}, \qquad (1)$$

where r denotes a locality radius or spatial resolution parameter.

This definition may be replaced by an equivalent spatial partitioning or indexing mechanism without loss of generality.

The local spatial structure is characterized by the tuple:

$$\mathcal{S}(x_i) = (\rho_i, \kappa_i, \gamma_i), \qquad (2)$$

where:

$-$ ρ_i denotes local point density,

$-$ κ_i denotes local spatial compactness,

$-$ γ_i denotes neighborhood coherence.

Such representation enables capturing spatial heterogeneity and scale-dependent spatial properties while remaining computationally tractable. For each point x_i, a local spatial neighborhood $\mathcal{N}(x_i)$is defined as a spatially bounded subset of X, constructed based on spatial proximity or spatial partitioning principles. The neighborhood definition is required to be independent of the clustering algorithm and consistent across different spatial scales.

The local spatial structure is modeled through the properties of these neighborhoods, capturing three essential aspects: local point density, spatial compactness, and neighborhood coherence. This representation allows the evaluation process to reflect spatial variability and local patterns that are typically obscured by global aggregation. By operating at the neighborhood level, the proposed model enables scale-aware analysis while maintaining computational tractability for large datasets.

Based on the modeled local spatial structure, cluster quality evaluation is decomposed into a set of local quality indicators computed independently within each spatial neighborhood. This decomposition enables the assessment of clustering quality at a granular level while avoiding exhaustive pairwise distance computations.

For a given neighborhood $\mathcal{N}(x_i)$, local quality indicators are defined to quantify the following properties:

- *Local compactness*, reflecting the spatial concentration of points belonging to the same cluster within the neighborhood;

- *Density consistency*, characterizing the agreement between local point density and the assigned cluster structure;

- *Spatial coherence*, measuring the spatial continuity of cluster assignments within the neighborhood.

Let $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$ denote a clustering result applied to dataset X. For each neighborhood $\mathcal{N}(x_i)$, a set of local quality indicators is computed.

Local compactness is defined as the inverse of the average intra-cluster distance within the neighborhood:

$$\kappa_i = \left(\frac{1}{| \mathcal{N}(x_i) |} \sum_{x_j \in \mathcal{N}(x_i)} d\left(x_j, \mu_{c(j)}\right) \right)^{-1}, \qquad (3)$$

where $\mu_{c(j)}$ denotes the centroid of the cluster to which point $x_j$belongs.

Local density consistency measures the agreement between local point density and cluster membership:

$$\rho_i = \frac{| \mathcal{N}(x_i) |}{\pi r^2} \qquad (4)$$

$$\delta_i = | \rho_i - \bar{\rho}_{c(i)} | \qquad (5)$$

where $\bar{\rho}_{c(i)}$ is the average density of cluster $c(i)$. Lower values of δ_i indicate higher consistency.

Spatial coherence is defined as the proportion of neighborhood points sharing the same cluster label:

$$\gamma_i = \frac{1}{|\mathcal{N}(x_i)|} \sum_{x_j \in \mathcal{N}(x_i)} \mathbb{I}(c(x_j) = c(x_i)) \qquad (6)$$

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

Each local indicator is computed using only neighborhood-level information and cluster membership relations, making the approach applicable to crisp, fuzzy, and hierarchical clustering results. Importantly, these indicators are formulated independently of the specific clustering algorithm, ensuring algorithmic generality. The use of local indicators enables robust evaluation in regions with varying density and supports meaningful interpretation of cluster quality at different spatial scales.

To obtain an overall assessment of clustering quality, the locally computed indicators are aggregated into a global quality measure. Let q_j denote the local quality score associated with neighborhood j. The global cluster quality evaluation function Q is defined as an aggregation of the set $\{q_j\}$, subject to normalization and weighting constraints.

Let q_i denote the composite local quality score for neighborhood $\mathcal{N}(x_i)$, defined as:

$$q_i = \alpha\kappa_i + \beta(1 - \delta_i) + \gamma\gamma_i \qquad (7)$$

where $\alpha, \beta, \gamma \geq 0$ and $\alpha + \beta + \gamma = 1$.

The global cluster quality evaluation score Q is obtained via normalized aggregation:

$$Q = \frac{1}{n} \sum_{i=1}^{n} q_i \qquad (8)$$

The aggregation strategy is designed to satisfy three key requirements: scalability, interpretability, and scale consistency. First, aggregation is performed using linear or weighted combination schemes that avoid pairwise dependencies between neighborhoods. Second, normalization ensures comparability of local indicators across regions with differing spatial densities. Third, the aggregation process preserves relative quality differences across spatial scales, preventing dominance of either dense or sparse regions.

This aggregation framework allows local spatial characteristics to contribute proportionally to the global quality assessment, resulting in a stable and interpretable evaluation measure. By explicitly integrating locality-driven indicators into a unified global score, the proposed strategy provides a principled and scalable alternative to classical cluster validation metrics for large-scale geospatial data.

To ensure a consistent and scalable definition of local spatial neighborhoods, the proposed evaluation framework employs hexagonal spatial indexing based on the H3 system. H3 provides a hierarchical discretization of the Earth's surface into hexagonal cells with predefined resolutions, enabling uniform spatial partitioning and efficient neighborhood traversal. An example of geospatial coverage using H3 hexagonal cells at a fixed resolution is illustrated in Fig. 1.

Let $H_r = \{h_1, h_2, \ldots, h_m\}$ denote the set of H3 cells at resolution level r. Each geospatial point $x_i \in X$ is mapped to a corresponding hexagonal cell $h(x_i) \in H_r$ according to its spatial coordinates. The local spatial neighborhood of a cell h is defined as:

$$\mathcal{N}(h) = \{ h_j \in H_r \mid d_H(h, h_j) \leq k \}, \qquad (10)$$

where $d_H(\cdot,\cdot)$ denotes the topological distance between hexagonal cells and k specifies the neighborhood radius in terms of adjacent rings.

As shown in Fig. 1, the hexagonal structure ensures compact and isotropic neighborhood boundaries, which reduces directional bias compared to square grid representations. The hierarchical nature of H3 further enables multi-scale analysis by varying the resolution level r, supporting scale-consistent cluster quality evaluation.

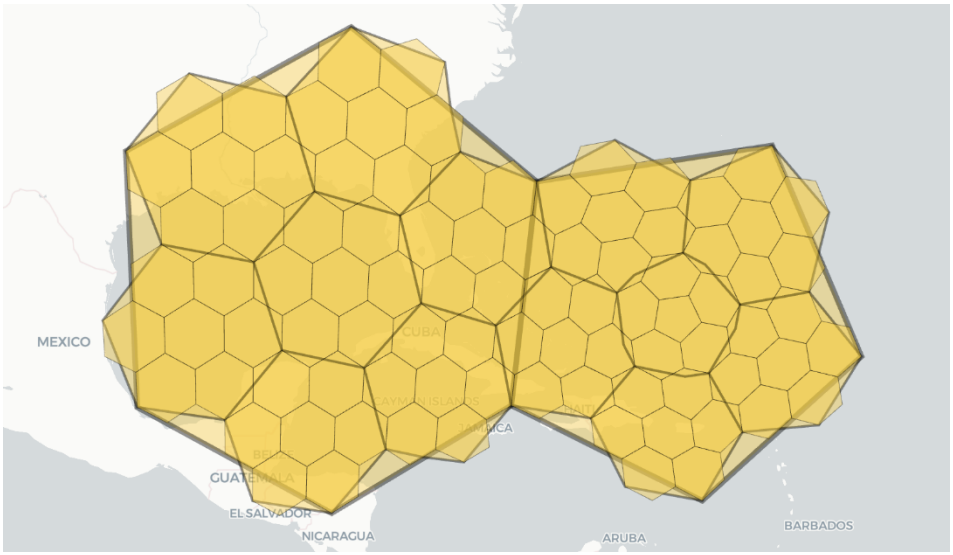

Fig.1. Example of hexagonal spatial partitioning using H3 indexing at a fixed resolution. Each hexagonal cell defines a local spatial neighborhood used for computing locality-driven cluster quality indicators.

Within this framework, local quality indicators are computed at the cell level by aggregating point-level information inside each hexagonal cell. This cell-based representation allows the proposed evaluation method to remain computationally efficient and spatially consistent, making it suitable for large-scale geospatial analytics and visualization tasks.

Based on the previously defined local spatial structure and quality indicators, the proposed evaluation framework follows a locality-driven aggregation strategy implemented through hexagonal spatial indexing. The complete evaluation procedure is formally specified in Fig.2, which summarizes all computational steps and their execution order.

As illustrated in Fig.2, at the first stage each geospatial data point is mapped to a hexagonal spatial cell using the H3 indexing scheme at a fixed resolution level r. This operation induces a spatial partition of the dataset into disjoint local regions. For each cell, a local neighborhood is constructed by retrieving adjacent cells within a predefined topological radius $k$using the H3 k-ring operator, resulting in a compact and isotropic spatial context.

At the second stage, clustering information is aggregated at the cell level. As shown in Fig.2 (lines 8–12), cluster memberships of all points assigned to the same hexagonal cell are summarized into compact cell-level representations. Subsequently, for each cell neighborhood, these summaries are combined to compute local quality indicators characterizing compactness, density consistency, and spatial coherence. Importantly, all indicators are evaluated exclusively using information from the cell and its neighborhood, thereby avoiding any global pairwise distance computations.

At the final stage, local quality scores are computed independently for each cell and normalized through global aggregation, as described in Fig.2 (lines 22–26). The resulting global cluster quality measure ensures proportional contribution of local regions while preserving robustness to spatial density heterogeneity and variations in spatial scale.

The local quality computation is embarrassingly parallel at the H3 cell level, as each neighborhood is evaluated independently. Synchronization is required only at the final aggregation stage, ensuring scalability for large-scale geospatial datasets.

Classical internal cluster validity indices commonly applied in clustering analysis rely on global or pairwise distance computations. For a dataset consisting of n spatial objects, such methods require evaluating distances between all pairs of points, which leads to quadratic time complexity $O(n^2)$ and quadratic memory requirements. This complexity severely limits their applicability to large-scale geospatial datasets and makes them unsuitable for interactive or near real-time analytical environments.

Algorithm 1 H3-Based Local Cluster Quality Evaluation (Parallel Cell Processing)

function GETNEIGHBORHOOD(h, k, H_r)
  $S \leftarrow$ H3KRing(h, k)
  return $S \cap H_r$
end function

function EVALUATECLUSTERQUALITY(X, C, r, k)
  $H_r \leftarrow \emptyset$
  $A \leftarrow$ empty map cell $\rightarrow$ membership summary
for each $x_i \in X$ **do**
    $h_i \leftarrow$ H3(x_i, r)
    $H_r \leftarrow H_r \cup \{h_i\}$
    update $A[h_i]$ with membership of x_i under C
end for
  $Q_{\text{local}} \leftarrow \emptyset$

for all cells $h \in H_r$ **in parallel do**
    $N(h) \leftarrow$ GETNEIGHBORHOOD(h, k, H_r)
    aggregate neighborhood memberships using $\{A[h_j] \mid h_j \in N(h)\}$
    compute local compactness κ_h
    compute density consistency δ_h
    compute spatial coherence γ_h
    $q_h \leftarrow \alpha\kappa_h + \beta(1 - \delta_h) + \gamma\gamma_h$
    atomic add q_h to Q_{local}
end for
synchronize all parallel tasks
$Q \leftarrow \frac{1}{|H_r|} \sum q_h \in Q_{\text{local}}$
return Q
end function

Fig.2. Algorithmic workflow of H3-based local cluster quality evaluation with parallel processing at the cell level.

The proposed evaluation framework is based on locality-driven computation enabled by hierarchical hexagonal spatial indexing. Each spatial object is mapped to a discrete spatial cell using H3 indexing in linear time $O(n)$. Let m denote the number of occupied hexagonal cells at a given resolution level, where $m \ll n$. Local neighborhoods are constructed at the cell level using a fixed topological radius, resulting in a bounded number of neighboring cells independent of n.

Local quality indicators are computed within each cell and its neighborhood, yielding a total computational complexity of $O(n + m)$, which reduces to $O(n)$ in practice. Unlike classical approaches, the proposed method avoids the construction of global distance matrices and does not require exhaustive pairwise comparisons. Memory consumption is therefore linear with respect to the dataset size.

Furthermore, the independence of local computations across spatial cells enables efficient parallelization. Local quality indicators can be evaluated independently for each cell or group of cells, making the proposed framework well suited for distributed and cloud-based processing of large-scale geospatial data.

Overall, the transition from quadratic $O(n^2)$ to linear $O(n)$ complexity constitutes a key scalability advantage of the proposed approach and underpins its applicability to modern large-scale geospatial analytics.

Experimental Setup and Datasets. The experimental evaluation was conducted using the publicly available Transportation Network Providers Trips 2025 dataset provided by the City of Chicago. The dataset contains large-scale geospatial records of ride-hailing trips, including spatial pickup and drop-off locations and associated temporal attributes. Owing to its size, spatial heterogeneity, and real-world origin,

this dataset is well suited for evaluating the scalability and robustness of cluster quality assessment methods under realistic conditions.

To analyze the impact of data volume on computational performance, multiple subsets of increasing size were extracted from the original dataset. The following dataset sizes were considered: $n = \{100, 10^3, 10^4, 5 \times 10^4, 10^5, 2 \times 10^5, 5 \times 10^5, 10^6, 5 \times 10^6\}$.
Each subset was generated by random sampling without replacement in order to preserve the original spatial distribution.

For clustering of spatial points, the Supercluster library (Mapbox) was employed. Supercluster implements a hierarchical spatial clustering algorithm optimized for large-scale point datasets and interactive web mapping applications. The algorithm builds a spatial index using a KDBush (KD-tree–based) structure and generates clusters across multiple zoom levels by aggregating points within a predefined spatial radius. This approach enables efficient preprocessing with approximately $O(n \log n)$ complexity and fast cluster retrieval at different spatial scales. The use of Supercluster ensures stable and scalable clustering results for large geospatial datasets, making it suitable for evaluating cluster quality metrics under realistic large-scale conditions.

The proposed locality-driven cluster quality evaluation method, optimized using H3 spatial indexing, was compared against a widely used internal cluster validity index the Silhouette coefficient, which evaluates clustering quality based on relative intra- and inter-cluster distances.

For each dataset size, clustering results were evaluated using the classical Silhouette coefficient and the proposed H3-based evaluation method. To ensure statistical stability of the obtained measurements, all experiments were performed over 100 independent runs per dataset size, and average execution times were used for comparison, reducing the influence of random fluctuations and runtime noise. All methods were executed under identical computational conditions. For the proposed approach, a fixed H3 resolution level and neighborhood radius were used across all experiments to ensure consistency.

The reported number of clusters (clusters count) corresponds to the number of spatial clusters produced by the Supercluster algorithm at a given zoom level. This value varies with both dataset size and spatial aggregation scale, reflecting the hierarchical nature of spatial clustering across different zoom levels.

The primary evaluation criterion was the computational time required to compute the cluster quality score as a function of dataset size. This experimental design enables a direct comparison of scalability properties, highlighting the contrast between classical evaluation methods relying on global or pairwise distance computations and the proposed locality-driven framework.

As summarized in Table 1, the runtime of the classical Silhouette coefficient grows rapidly with increasing dataset size. This behavior is consistent with the worst-case quadratic time complexity $O(n^2)$ of Silhouette evaluation based on pairwise distance computations. For datasets exceeding 10^5 spatial objects, the classical approach becomes impractical or infeasible due to excessive execution time, resulting in timeouts for large-scale scenarios.

Table 1. Runtime scalability and value consistency of classical and H3-based cluster quality indices

ata set Size	M ethod	Map zoom level 9			Map zoom level 11			Map zoom level 16		
		lusters count	alue	un time	R	lusters count	alue	un time	R	lusters count	alue	un time	R
00	S ilhouette		.0466	.1ms	0		.0000	.1ms	0		.0000	.1ms	0
1	H 3-based		.0466	.5ms	0		.0000	.5ms	0		.0000	.4ms	0
000	S ilhouette	30	.2608	.7ms	3	8	.0445	.0ms	1		.0000	.6ms	0
1	H 3-based	30	.2608	.3ms	4	8	.0097	.6ms	3		.0000	.5ms	3
0000	S ilhouette	625	.1783	31.8ms	1	695	.1891	92.5ms	1		.0002	.9ms	7
1	H 3-based	625	.1783	4.6ms	3	695	.1815	1.1ms	4		.0000	9.3ms	3
0000	S ilhouette	038	.1058	27.8ms	3	3703	.2519	.80s	6	2	.0027	5.0ms	9

	H			6			2			2
	3-based	038	.1058	1.9ms	3703	.2486	14.3ms	2	.0002	27.4ms
	S			3			1			4
00000	ilhouette	062	.1073	91.1ms	1675	.2057	9.18s	44	.0045	24.4ms
	H			6			3			5
1	3-based	062	.1072	9.3ms	1675	.2010	61.2ms	44	.0005	11.1ms
	S			4			3			2
00000	ilhouette	069	.1055	85.9ms	7356	.1518	9.08s	98	.0089	.75s
	H			7			5			1
2	3-based	069	.1055	4.4ms	7356	.1466	41.6ms	98	.0017	.14s
	S			6			6			3
00000	ilhouette	082	.1084	97.9ms	0901	.1149	6.83s	051	.0203	7.09s
	H			7			7			3
5	3-based	082	.1084	8.2ms	0901	.1095	38.2ms	051	.0086	.54s
	S			t			t			t
0,000	ilhouette	065		imeout (>=30m)	1901		imeout (>=30m)	3786		imeout (>=30m)
	H			7			8			7
1,00	3-based	065	.1064	6.8ms	1901	.0960	41.6ms	3786	.0250	.44s
	S			t			t			t
0,000	ilhouette	065		imeout (>=30m)	2235		imeout (>=30m)	10081		imeout (>=30m)
	H			4			2			8
5,00	3-based	065	.1118	30.9ms	2235	.0994	.72s	10081	.1294	4.61s

ф

Figure 2 clearly illustrates the contrasting scalability behavior of the two evaluation approaches. The runtime of the classical Silhouette coefficient increases steeply as the dataset size grows, reflecting its reliance on global pairwise distance computations and worst-case quadratic complexity. In contrast, the H3-based evaluation method scales much more smoothly, maintaining low execution times even for large datasets. This confirms that locality-driven spatial aggregation effectively mitigates the computational bottleneck inherent in classical cluster quality evaluation.

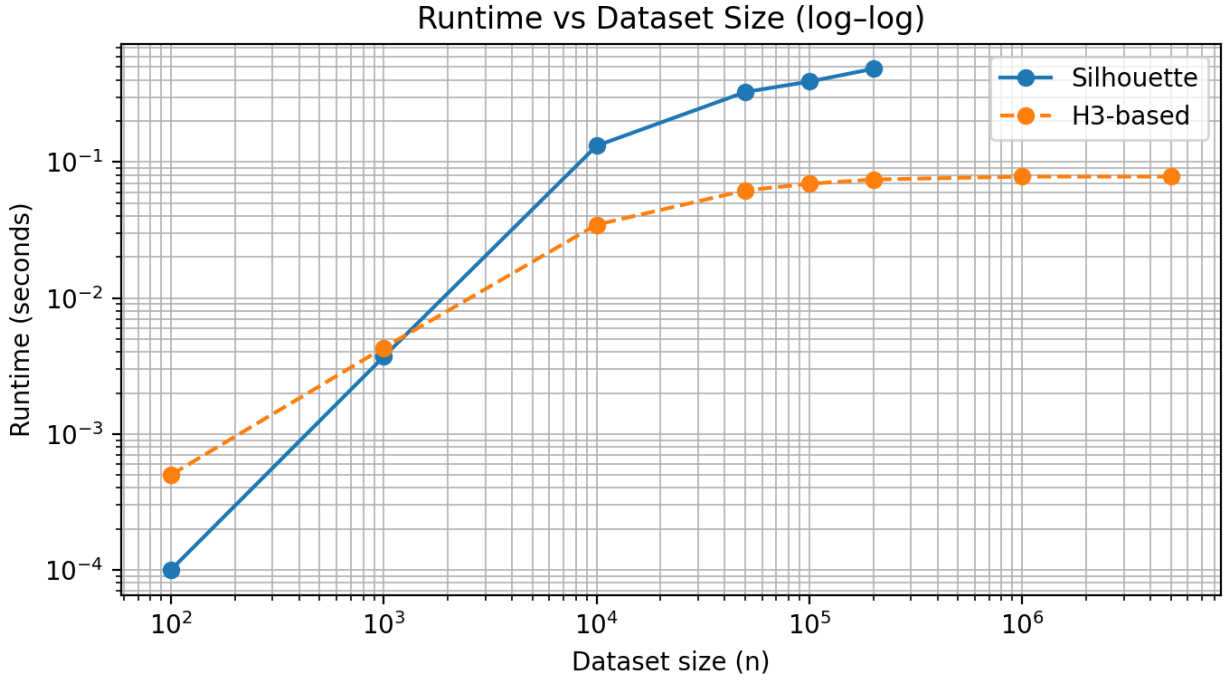

Fig.2. Runtime comparison of the classical Silhouette coefficient and the proposed H3-based evaluation method as a function of dataset size on a log–log scale. The classical approach exhibits rapidly increasing execution time with dataset growth, while the H3-based method demonstrates near-linear scalability across all evaluated data sizes.

As shown in Figure 3, the performance advantage of the H3-based approach increases consistently with dataset size. For small datasets, the speedup is limited, reflecting comparable overheads. However, as

the number of spatial objects grows, the speedup rapidly exceeds one order of magnitude, demonstrating that the proposed method becomes increasingly advantageous at scale. This trend highlights the suitability of the H3-based framework for large-scale and interactive geospatial analysis scenarios.

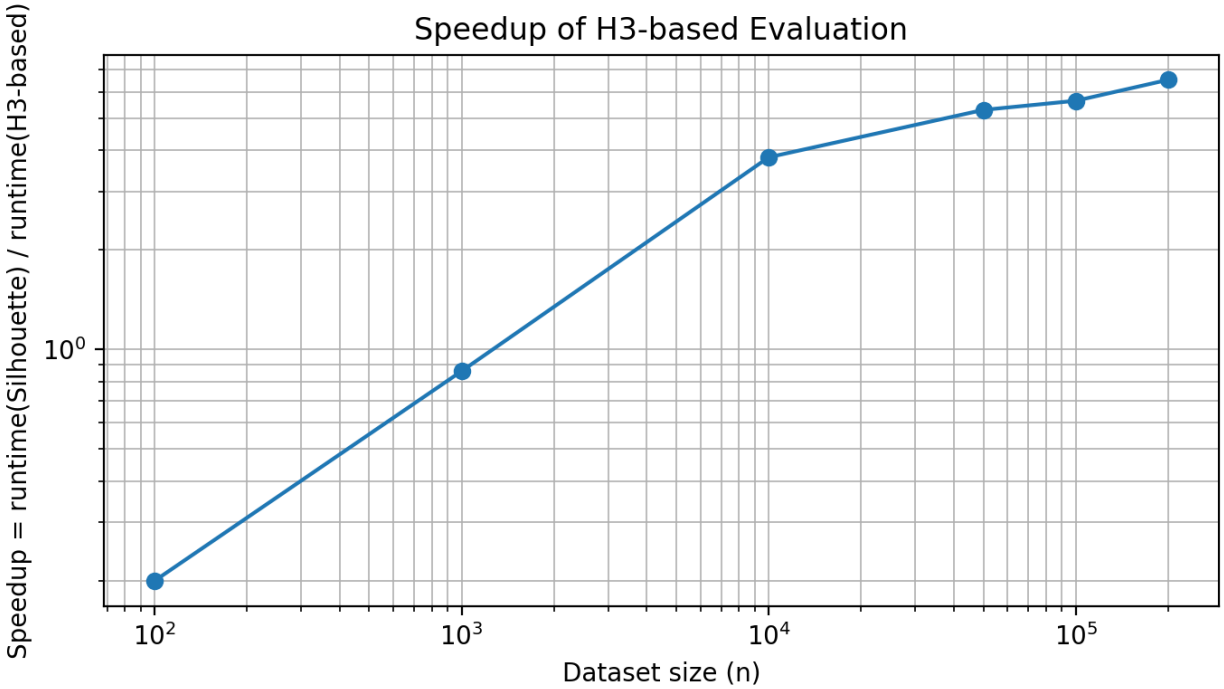

Fig. 3. Speedup factor of the proposed H3-based evaluation relative to the classical Silhouette coefficient as a function of dataset size. Speedup is defined as the ratio of classical runtime to H3-based runtime and is shown on a logarithmic scale.

Taken together, the presented results indicate that the proposed H3-based evaluation framework fundamentally changes the operational role of cluster quality assessment in geospatial analytics. Rather than serving as an offline diagnostic step applied only to small or subsampled datasets, cluster quality evaluation becomes a computationally affordable component that can be integrated into large-scale processing pipelines, iterative model selection, and interactive exploration workflows. This shift enables practitioners to assess clustering outcomes directly on full-resolution data and across multiple spatial scales, without resorting to approximation through data reduction. Consequently, the framework expands the practical applicability of established cluster validity indices to modern geospatial scenarios characterized by high data volume, spatial heterogeneity, and real-time analytical requirements.

Conclusions. This study addressed the challenge of cluster quality evaluation for large-scale geospatial data by proposing a locality-driven evaluation framework optimized through hierarchical hexagonal spatial indexing. The results demonstrate that classical cluster validity indices, such as the Silhouette coefficient, can be efficiently reformulated to operate on local spatial neighborhoods without compromising evaluation semantics. Experimental analysis shows that the proposed H3-based approach achieves linear-time scalability, enabling cluster quality evaluation for datasets containing millions of spatial objects, which is impractical for traditional pairwise distance-based methods. At the same time, numerical consistency with classical evaluation values is preserved, with only limited and interpretable deviations attributable to spatial aggregation effects. Overall, the proposed framework transforms cluster quality evaluation from a computational bottleneck into a scalable analytical component, supporting practical large-scale geospatial clustering and interactive spatial analysis.

Contributions:

- A locality-driven paradigm for cluster quality evaluation is proposed, shifting evaluation from global pairwise distance computations to neighborhood-based spatial analysis while preserving the semantics of classical validity indices.
- A scalable H3-based evaluation framework is developed, enabling linear-time computation of cluster quality metrics for large-scale geospatial datasets through hierarchical hexagonal spatial indexing and bounded neighborhood aggregation.

- An algorithm-agnostic optimization strategy is introduced, allowing classical cluster validity indices, including the Silhouette coefficient, to be applied efficiently to crisp, fuzzy, and hierarchical clustering results.

- An empirical analysis of scalability and value consistency is provided, demonstrating that the proposed approach achieves orders-of-magnitude runtime improvements while maintaining numerical stability and interpretability of cluster quality values.

References

1. Altieri, F., Pietracaprina, A., Pucci, G., & Vandin, F. (2021). Scalable distributed approximation of internal measures for clustering evaluation. In Proceedings of the 2021 SIAM International Conference on Data Mining (SDM) (pp. 648–656). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611976700.73
2. Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999, June). OPTICS. Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. SIGMOD/PODS99: International Conference on Management of Data and Symposium on Principles of Database Systems, Philadelphia Pennsylvania USA. https://doi.org/10.1145/304182.304187
3. Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical Analysis, 27(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x
4. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1), 243–256. https://doi.org/10.1016/j.patcog.2012.07.021
5. AssunÇão, R. M., Neves, M. C., Câmara, G., & Da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. International Journal of Geographical Information Science, 20(7), 797–811. https://doi.org/10.1080/13658810600665111
6. Ben Ncir, C.-E., Hamza, A., & Bouaguel, W. (2021). Parallel and scalable Dunn Index for the validation of big data clusters. Parallel Computing, 102(102751), 102751. https://doi.org/10.1016/j.parco.2021.102751
7. Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics: Theory and Methods, 3(1), 1–27. https://doi.org/10.1080/03610927408827101
8. Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Advances in Knowledge Discovery and Data Mining (pp. 160–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14
9. Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224–227. https://www.ncbi.nlm.nih.gov/pubmed/21868852
10. Dunn†, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4(1), 95–104. https://doi.org/10.1080/01969727408546059
11. Fleischer, C. E. (2021). Using the max-p regions problem algorithm to define regions for energy system modelling. MethodsX, 8(101211), 101211. https://doi.org/10.1016/j.mex.2021.101211
12. Gaido, M. (2023). Distributed Silhouette algorithm: Evaluating clustering on big data. In arXiv [cs.DC]. arXiv. http://arxiv.org/abs/2303.14102
13. Getis, A., & Ord, J. K. (2010). The analysis of spatial association by use of distance statistics. In Advances in Spatial Science (pp. 127–145). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-01976-0_10
14. Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). International Journal of Geographical Information Science, 22(7), 801–823. https://doi.org/10.1080/13658810701674970
15. Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems.
16. Halkidi, M., & Vazirgiannis, M. (2002). Clustering validity assessment: finding the optimal partitioning of a data set. Proceedings 2001 IEEE International Conference on Data Mining. 2001 IEEE International Conference on Data Mining, San Jose, CA, USA. https://doi.org/10.1109/icdm.2001.989517
17. Jeon, H., Aupetit, M., Shin, D., Cho, A., Park, S., & Seo, J. (2025). Measuring the validity of clustering validation datasets. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(6), 5045–5058. https://doi.org/10.1109/TPAMI.2025.3548011
18. Kulldorff, M. (1997). A spatial scan statistic. Communications in Statistics: Theory and Methods, 26(6), 1481–1496. https://doi.org/10.1080/03610929708831995
19. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., & Wu, S. (2013). Understanding and enhancement of internal clustering validation measures. IEEE Transactions on Cybernetics, 43(3), 982–994. https://doi.org/10.1109/TSMCB.2012.2220543
20. Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical Navigable Small World graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(4), 824–836. https://doi.org/10.1109/TPAMI.2018.2889473
21. Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1–2), 17–23. https://doi.org/10.1093/biomet/37.1-2.17
22. Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., & Sander, J. (2014, April 28). Density-based clustering validation. Proceedings of the 2014 SIAM International Conference on Data Mining. Proceedings of the 2014 SIAM International Conference on Data Mining. https://doi.org/10.1137/1.9781611973440.96
23. Nelson, J. K. (2025). Difference mapping approach to detecting the cartographic effects of the modifiable areal unit problem. KN - Journal of Cartography and Geographic Information. https://doi.org/10.1007/s42489-025-00199-9
24. Nguyen, L. T. T., Nguyen, T. T. D., Bui, Q.-T., & Vo, B. (2025). Geospatial data clustering in network space: A survey. Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 15(2). https://doi.org/10.1002/widm.70023

25. Ram, A., Jalal, S., Jalal, A. S., & Kumar, M. (2010). A density based algorithm for discovering density varied clusters in large spatial databases. International Journal of Computer Applications, 3(6), 1–4. https://doi.org/10.5120/739-1038
26. Randriamihamison, N., Vialaneix, N., & Neuvial, P. (2021). Applicability and interpretability of ward's hierarchical agglomerative clustering with or without contiguity constraints. Journal of Classification, 38(2), 363–389. https://doi.org/10.1007/s00357-020-09377-y
27. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
28. Sarle, W. S., Jain, A. K., & Dubes, R. C. (1990). Algorithms for Clustering Data. Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences, 32(2), 227. https://doi.org/10.2307/1268876
29. Ullmann, T., Hennig, C., & Boulesteix, A.-L. (2022). Validation of cluster analysis results on validation data: A systematic framework. Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 12(3). https://doi.org/10.1002/widm.1444
30. Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(8), 841–847. https://doi.org/10.1109/34.85677

Історія статті:
Отримано: 13.02.2026 Доопрацьовано: 01.03.2026 Прийнято до друку: 23.03.2026 Опубліковано: 29.03.2026