CEUR-WS.org/Vol-4141/paper10.pdf

A benchmark methodology for urban traffic pattern clustering using SUMO-based expert-verified ground truth

Vitaliy Pavlyshyn[1,*,†], Eduard Manziuk[1,†], Adnène Arbi[2,†], Nebojsa Bacanin[3,†] and Iurii Krak[4,5,†]

[1]*Khmelnytskyi National University, 11, Instytuts'ka str., 29016 Khmelnytskyi, Ukraine*
[2]*University of Carthage, Avenue de la République, 1054 Amilcar, Tunis, Tunisia*
[3]*Singidunum University, 32 Danijelova St., 11000 Belgrade, Serbia*
[4]*Taras Shevchenko National University of Kyiv, 64/13, Volodymyrska str., Kyiv, 01601, Ukraine*
[5]*Glushkov Cybernetics Institute, 40, Glushkov Ave., Kyiv, 03187, Ukraine*

Abstract

Identifying urban traffic patterns is critical for reducing CO_2 emissions, yet existing research lacks standardized benchmarks for objectively evaluating clustering algorithms. This fundamental gap prevents accurate assessment because real-world traffic data typically lacks ground truth labels, making the validation of clustering quality impossible. In this work, we propose a methodology for controlled comparison of clustering algorithms using expert-verified ground truth labels derived from SUMO simulations of real urban scenarios. We systematically evaluate six clustering algorithms (HDBSCAN, K-Means, MeanShift, AffinityPropagation, BayesianGMM, AgglomerativeClustering) on both aggregated and concatenated vector representations of traffic data. Our experiments reveal that HDBSCAN achieves the highest accuracy in recovering ground truth scenarios (ARI=0.73, V-measure=0.79) on aggregated data, outperforming K-Means by 0.03 in ARI. Furthermore, aggregated representations systematically outperformed detailed temporal data for all algorithms with an average ARI improvement of 0.15. The study provides a validated benchmarking methodology enabling objective algorithm selection for traffic management systems aimed at emission reduction.

Keywords

Clustering, traffic patterns, SUMO, urban traffic, traffic management, CO_2 emissions

1. Introduction

The growth of urbanization and urban traffic intensity creates serious challenges for sustainable city development, especially in the context of combating climate change. The transport sector accounts for over one-third of CO_2 emissions from final energy consumption in cities, making traffic flow optimization critically important for achieving climate goals [1]. Identifying characteristic urban traffic patterns enables the development of traffic management strategies and reduction of environmental impact [2].

Modern urban transport systems face a critical problem of lacking reliable methods for objectively assessing the quality of clustering algorithms when analyzing traffic flows. Most existing research is based on real GPS data or traffic detector readings, which by their nature lack ground truth labels, making accurate assessment of clustering quality impossible. This fundamental problem creates a significant barrier to developing traffic management systems aimed at reducing CO_2 emissions.

The previously unsolved part of the general problem of traffic flow optimization lies in the absence of standardized methodologies for controlled comparison of clustering algorithms under conditions where the true structure of traffic patterns is known. This gap is especially critical in the context of

ExplAI-2025: Advanced AI in Explainability and Ethics for the Sustainable Development Goals, November 07, 2025, Khmelnytskyi, Ukraine

*Corresponding author.

†These authors contributed equally.

Vitaliy@ualeaders.com (V. Pavlyshyn); eduard.em.km@gmail.com (E. Manziuk); adnene.arbi@insat.ucar.tn (A. Arbi); nbacanin@singidunum.ac.rs (N. Bacanin); iurii.krak@knu.ua (I. Krak)
0009-0003-0702-9359 (V. Pavlyshyn); 0000-0002-7310-2126 (E. Manziuk); 0000-0002-2834-0248 (A. Arbi); 0000-0002-2062-924X (N. Bacanin); 0000-0002-8043-0785 (I. Krak)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-22

cities' climate commitments, where accurate identification of traffic patterns can significantly impact emission reduction.

Manifestations of this problem include the inability to determine which clustering algorithm best identifies real traffic patterns under different urban conditions, lack of consensus on optimal metrics for evaluating traffic data clustering quality, and shortage of controlled experimental conditions for validating research results in this field.

The main contribution of this research is a proposed methodology for controlled comparison of clustering algorithms for traffic data using expert-verified ground truth labels created from real urban traffic scenarios, which allows objective evaluation of different clustering approaches under conditions maximally approximating real transport systems. The research also contributes to understanding the impact of different traffic data aggregation approaches on clustering quality, which has direct practical significance for developing traffic management systems oriented toward reducing CO_2 emissions through traffic flow optimization.

The structure of the paper is as follows. The "Literature Review" section analyzes existing approaches to traffic data clustering. The "Materials and Methods" section describes the experimental methodology and algorithms. The "Results" section presents quantitative algorithm indicators. The "Discussion" section interprets the obtained results and compares them with existing approaches.

2. Related works

This section provides an overview of current research on traffic flow clustering, simulation approaches using SUMO for CO_2 emission assessment, and intelligent traffic management systems. Traffic data clustering represents an actively developing research area evolving under the influence of growing needs to reduce CO_2 emissions from transport. The transport sector accounts for over one-third of CO_2 emissions from final consumption [1], making traffic flow optimization critically important for achieving climate goals.

Analysis of current research reveals two dominant approaches: centroid-based and density-based clustering methods. Systematic analysis of K-means application for zone classification by congestion level showed its quality in identifying delay patterns associated with different types of traffic flows [2]. However, limitations of centroid methods stimulated development of hybrid approaches. Combining pairwise comparison with density-based methods proved applicable for processing multidimensional time series, demonstrating advantages over traditional centroid algorithms [3].

Mathematical models of urban mobility optimization are developing in parallel [4], integrating with graph-based approaches for analyzing spatiotemporal cluster evolution [5]. These methodologies allow not only identifying static patterns but also tracking their dynamics over time, which is critical for traffic flow forecasting.

Comparative algorithm analysis reveals HDBSCAN's advantage due to its ability to automatically determine the number of clusters and handle noise. The integration of visual analytics approaches with machine learning algorithms [6] provides methodological foundation for combining expert knowledge with automated pattern detection in complex datasets. Two-phase approaches integrating GIS and HDBSCAN demonstrated advantages in spatial analysis of accident-prone areas [7], confirmed by enhanced versions for multi-level spatial pattern analysis [8]. A critical advantage of HDBSCAN is detecting variable-density clusters, corresponding to real characteristics of urban traffic flows.

Further methodology development led to creation of emission-sensitive clustering algorithms [9], which extend dynamic pattern detection capabilities. For high-dimensional data, stratified density algorithms are proposed [10], solving the curse of dimensionality problem in big data.

The transition from real data to controlled experiments determines the growing role of simulation tools. SUMO became the standard thanks to integration capabilities with real sensor data [11]. Validation studies on heterogeneous transport conditions confirmed SUMO's universality through achieving high accuracy correspondence between simulation and real data [12]. Research on generating and calibrating microscopic urban models for different scenarios [13] provides methodological foundation for using

SUMO in transport system research.

General environmental emission reduction trends drive integration of traffic pattern analysis with emission assessment. Systematization of carbon emission reduction technologies [14] and development of multimodal approaches [15] demonstrate alignment of traffic optimization with climate goals. CO_2 emission forecasting using deep learning and explainable artificial intelligence achieved quite high accuracy [16], revealing that fuel consumption conditions in urban and suburban settings have greater impact than vehicle engine characteristics.

Predictive models for intersections using portable measurement systems and density clustering algorithms [17] provide detailed micro-level analysis, complementing macroscopic approaches. Combining big data and artificial intelligence opens new management possibilities. Adaptive traffic light control can reduce travel time by 11% during peak hours, extrapolating to annual CO_2 emission reduction of 31.73 million tons [18]. Recent work on AI-driven traffic signal control systems [19] demonstrates the direct applicability of machine learning approaches to emission reduction, reinforcing the practical importance of accurate traffic pattern identification for environmental objectives.

Modeling approaches for intelligent transport systems [20] emphasize the need for environmental considerations in urban mobility optimization, aligning with the emission reduction focus of this research. Multi-agent deep reinforcement learning approaches [21] and connected vehicle coordination systems [22] demonstrate evolution from centralized to distributed control. Cooperative traffic light control methods with deep learning [23] ensure coordination between multiple intersections, creating an adaptive control network.

Methodological analysis shows transition from one-dimensional to multi-level approaches. Comprehensive reviews emphasize the importance of quality simulation data [24], implemented through two-level machine learning architectures [25]. Integration of spatiotemporal data with real-time route optimization [26] and multi-scale models for medium-term forecasting [27] demonstrate growing complexity of predictive systems. Network approach to mobility analysis through cluster detection methods [28] is complemented by high-resolution cellular network data analysis [29]. Systematization of pattern identification methods using smart card data and deep learning application for spatiotemporal analysis [30] demonstrate evolution from descriptive to predictive urban mobility modeling.

Critical analysis of traditional methods reveals their limitations in determining optimal cluster numbers. Metaheuristic approaches [31] and evolutionary K-means methods [32] offer solutions through automatic parameter optimization, especially important for traffic data with unknown cluster structure. Diversity-based approaches to clustering [33] demonstrate that ensemble methods leveraging multiple clustering perspectives can improve robustness and accuracy, particularly relevant for heterogeneous traffic pattern identification. Selection criteria for ensemble models [34] provide theoretical basis for comparing multiple clustering algorithms systematically, which motivates the multi-algorithm evaluation approach adopted in this study.

Thus, analysis revealed that despite significant progress in the considered field, critical gaps remain. Absence of benchmark data complicates objective algorithm comparison. Fuzzy clustering application emphasizes the need for controlled experimental conditions and ground truth labels for reliable result validation. The conducted analysis reveals a fundamental contradiction: despite the diversity of available clustering algorithms and their theoretical advantages, absence of labeled benchmark datasets makes objective quality comparison impossible for traffic data. This contradiction determines the research goal: improving traffic pattern identification quality through developing an approach that ensures objective clustering algorithm comparison on controlled simulation data with known ground truth structure.

To achieve this goal, the following tasks are formulated:

1. Create a traffic flow simulation model verified by experts and based on real urban scenarios.
2. Conduct systematic comparison of six representative clustering algorithms using standardized metrics.

3. Materials and methods

3.1. General approach schema

The general schema of the proposed approach is shown in Figure 1. The proposed schema implements the research hypothesis that using a city map and empirical knowledge about existing traffic flow behavior, SUMO can simulate traffic movement that corresponds to real traffic and allows objective evaluation of clustering algorithm quality.

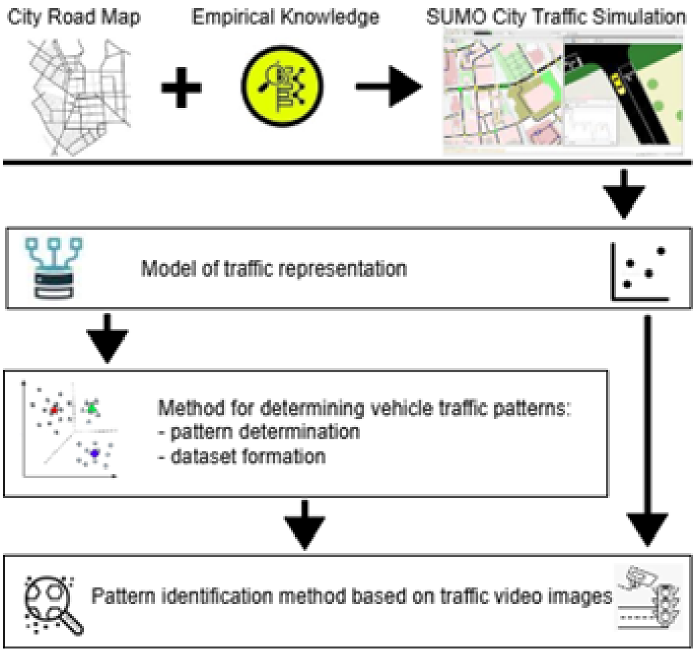
City Road Map
Empirical Knowledge
SUMO City Traffic Simulation
Model of traffic representation
Method for determining vehicle traffic patterns:
- pattern determination
- dataset formation
Pattern identification method based on traffic video images

Figure 1: General schema of the approach to comparing traffic pattern clustering algorithms.

The approach consists of five main stages: (1) creating a simulation model based on real urban scenarios; (2) generating traffic data in time window format; (3) converting data into vector representation for clustering; (4) applying clustering algorithms; (5) evaluating result quality using standardized metrics.

3.2. Traffic representation model

Traffic flow is formalized as a sequence of traffic light intersection states at discrete time moments. Each traffic light state s_j at time t_i is characterized by a vector of vehicle counts on each traffic lane. Two types of vector data representation are used for analysis: concatenated and averaged values.

3.2.1. Concatenated values

Concatenated values represent detailed temporal representation of traffic flow with preservation of complete information about vehicle count changes over time. For each 30-minute time window W, a vector V^{concat} of dimensionality $d = 70 \times 180 = 12,600$ is formed, where 70 is the total number of traffic lanes at all traffic lights, and 180 is the number of time slices (data is recorded every 10 seconds during 30 minutes):

$$V^{concat} = [t_1 s_1 l_1, t_1 s_1 l_2, \ldots, t_1 s_1 l_7, t_1 s_2 l_1, \ldots, t_i s_j l_k], \tag{1}$$

where $t_i s_j l_k$ is the number of vehicles on the k-th lane of the j-th traffic light at time moment t_i, $i \in \{1, 2, \ldots, 180\}$, $j \in \{1, 2, \ldots, 10\}$, $k \in \{1, 2, \ldots, m_j\}$, where m_j is the number of lanes at the j-th traffic light. This approach preserves temporal traffic dynamics but creates a high-dimensional feature space, which may lead to the curse of dimensionality problem in clustering.

3.2.2. Averaged values

Averaged values represent aggregated temporal representation, where one summary value is computed for each traffic lane over the entire window period. A vector V^{avg} of dimensionality $d = 70$ is formed:

$$V^{avg} = [s_1 l_1, s_1 l_2, \ldots, s_1 l_7, s_2 l_1, \ldots, s_j l_k], \tag{2}$$

where each component is computed as arithmetic mean:

$$s_j l_k = \frac{1}{T} \sum_{i=1}^{T} t_i s_j l_k, \tag{3}$$

where $T = 180$ is the number of time slices in the window. Thus, $s_j l_k$ is the average number of vehicles on the k-th lane of the j-th traffic light over the entire window period.

This approach sacrifices temporal detail for reducing feature space dimensionality by 180 times, providing better conditions for clustering algorithms and increasing resistance to short-term data fluctuations. The trade-off between information completeness and clustering quality is investigated experimentally by comparing results on both representation types.

3.3. Traffic pattern determination method

Traffic pattern identification is implemented through a sequential process integrating expert knowledge with controlled simulation to ensure objective clustering algorithm evaluation. The complete algorithmic procedure is presented in Algorithm 1 and Figure 2.

3.3.1. Creating base traffic scenarios

Detailed schema of traffic pattern determination method is shown in Figure 2. At the first stage, four base traffic scenarios are formed based on surveillance camera data analysis and expert knowledge from municipal traffic specialists. The morning scenario is characterized by intensive traffic to the city center and market zone, reflecting typical commuting migrations on working days. The evening scenario represents reverse flow from center and market to residential areas. The random scenario models uniformly distributed traffic without clearly expressed dominant direction. The special scenario reflects characteristic traffic from the peripheral Hrechany district, which differs from general city patterns due to its location specifics and transport infrastructure. Each scenario is verified by experts to ensure correspondence with real city traffic flows.

3.3.2. Traffic flow simulation

At the second stage, created scenarios are implemented in SUMO simulation environment version 1.15.0 using a real city map. The simulation covers an 11-hour period with recording states of 10 key traffic light intersections. Data is collected at 10-second intervals, providing sufficient temporal resolution for capturing traffic flow dynamics. In total, 4,080 traffic light state records are generated. The simulation is configured considering real urban road network parameters.

3.3.3. Time window formation

At the third stage, generated data is segmented into 30-minute time windows to ensure sufficient information volume for statistical pattern analysis. An overlapping window method with 10-minute shift step is used, allowing increase of observation count from 22 to 66 and ensuring temporal result stability.

3.3.4. Traffic data vectorization

At the fourth stage, each time window is converted into vector representation according to the formalization described in Section 3.2. Obtained vectors are standardized using z-score normalization.

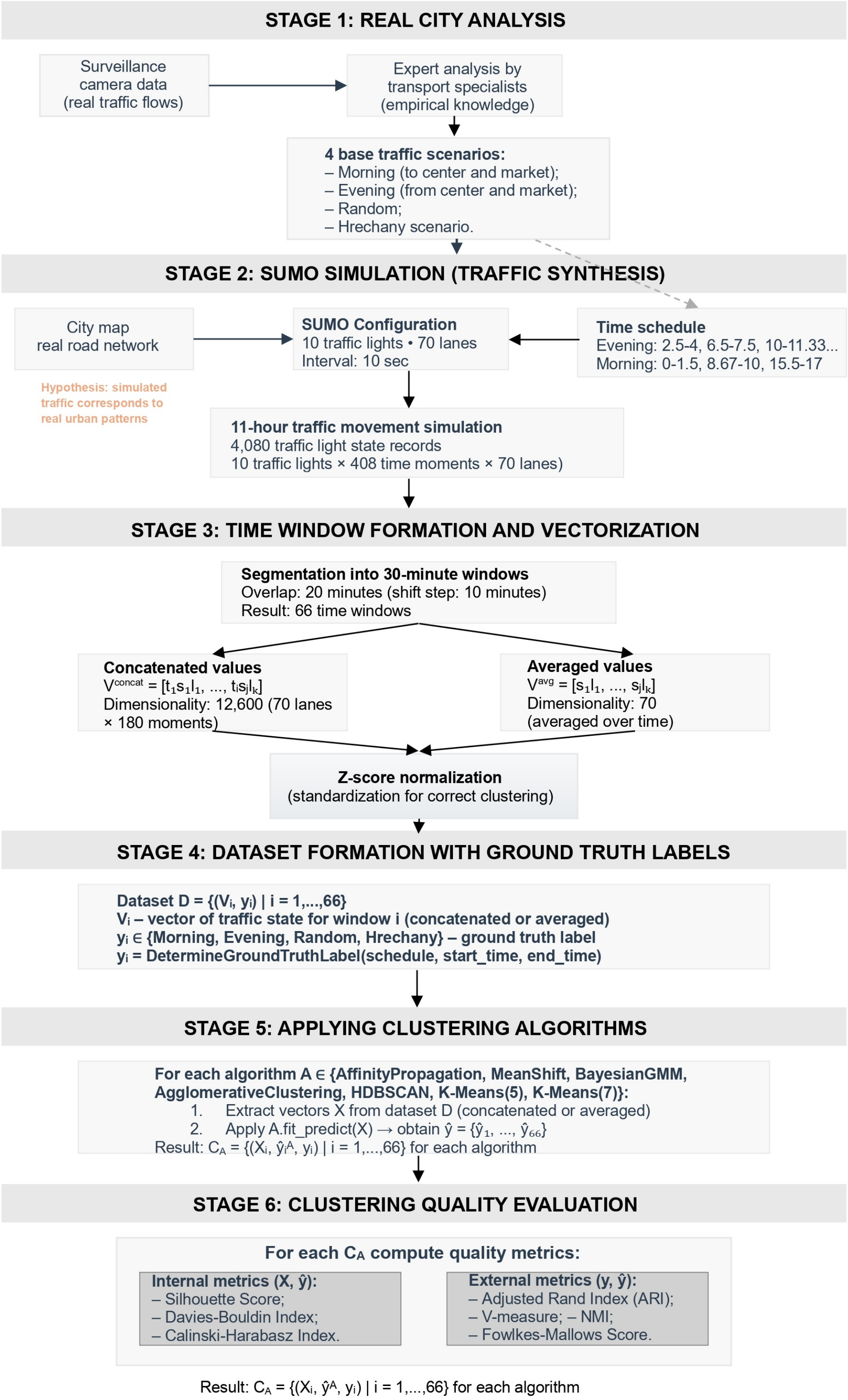
STAGE 1: REAL CITY ANALYSIS
Surveillance camera data (real traffic flows)
Expert analysis by transport specialists (empirical knowledge)
4 base traffic scenarios:
– Morning (to center and market);
– Evening (from center and market);
– Random;
– Hrechany scenario.
STAGE 2: SUMO SIMULATION (TRAFFIC SYNTHESIS)
City map real road network
SUMO Configuration
10 traffic lights • 70 lanes
Interval: 10 sec
Time schedule
Evening: 2.5-4, 6.5-7.5, 10-11.33...
Morning: 0-1.5, 8.67-10, 15.5-17
Hypothesis: simulated traffic corresponds to real urban patterns
11-hour traffic movement simulation
4,080 traffic light state records
10 traffic lights × 408 time moments × 70 lanes)
STAGE 3: TIME WINDOW FORMATION AND VECTORIZATION
Segmentation into 30-minute windows
Overlap: 20 minutes (shift step: 10 minutes)
Result: 66 time windows
Concatenated values
$V^{concat} = [t_1 s_1 l_1, ..., t_i s_j l_k]$
Dimensionality: 12,600 (70 lanes × 180 moments)
Averaged values
$V^{avg} = [s_1 l_1, ..., s_j l_k]$
Dimensionality: 70
(averaged over time)
Z-score normalization
(standardization for correct clustering)
STAGE 4: DATASET FORMATION WITH GROUND TRUTH LABELS
Dataset $D = \{(V_i, y_i) \mid i = 1,...,66\}$
V_i – vector of traffic state for window i (concatenated or averaged)
$y_i \in$ {Morning, Evening, Random, Hrechany} – ground truth label
y_i = DetermineGroundTruthLabel(schedule, start_time, end_time)
STAGE 5: APPLYING CLUSTERING ALGORITHMS
For each algorithm A ∈ {AffinityPropagation, MeanShift, BayesianGMM, AgglomerativeClustering, HDBSCAN, K-Means(5), K-Means(7)}:
1. Extract vectors X from dataset D (concatenated or averaged)
2. Apply A.fit_predict(X) → obtain $\hat{y} = \{\hat{y}_1, ..., \hat{y}_{66}\}$
Result: $C_A = \{(X_i, \hat{y}_i^A, y_i) \mid i = 1,...,66\}$ for each algorithm
STAGE 6: CLUSTERING QUALITY EVALUATION
For each C_A compute quality metrics:
Internal metrics (X, ŷ):
– Silhouette Score;
– Davies-Bouldin Index;
– Calinski-Harabasz Index.
External metrics (y, ŷ):
– Adjusted Rand Index (ARI);
– V-measure; – NMI;
– Fowlkes-Mallows Score.
Result: $C_A = \{(X_i, \hat{y}^A, y_i) \mid i = 1,...,66\}$ for each algorithm

Figure 2: Detailed schema of traffic pattern determination method.

3.3.5. Applying clustering algorithms

At the fifth stage, six representative clustering algorithms with optimized parameters determined through preliminary validation on pilot dataset are applied to vectorized data. AffinityPropagation is used with damping parameter equal to 0.8. MeanShift is applied with automatic bandwidth de-

Algorithm 1 Traffic Pattern Determination and Evaluation Method

```
Require: City map $M$, expert knowledge $E$, set of algorithms $A$
Ensure: Dataset $D$ with ground truth labels, evaluation results for each algorithm
// STAGE 1: Real city analysis and scenario creation
$camera\_data \leftarrow$ GetCameraObservations$(M)$
$S \leftarrow \{s_1, s_2, s_3, s_4\} \leftarrow$ ExpertAnalysis$(camera\_data, E)$
// $s_1$: Morning, $s_2$: Evening, $s_3$: Random, $s_4$: Hrechany
// STAGE 2: Traffic synthesis in SUMO (hypothesis implementation)
$sumo\_config \leftarrow$ ConfigureSUMO$(M, num\_lights = 10, interval = 10sec)$
$schedule \leftarrow$ CreateScenarioSchedule$(S, duration = 11hours)$
// Scenario distribution across simulation timeline
$traffic\_states \leftarrow []$
for $t \leftarrow 0$ to $39,600$ step 10 do                ▷ 11 hours × 3,600 sec/hour
    $current\_scenario \leftarrow schedule.$GetScenario$(t)$
    $state \leftarrow$ SUMO.Simulate$(current\_scenario, t)$
    // state: vector of vehicle counts on 70 lanes
    $traffic\_states.$append$(state)$
// STAGE 3: Time window formation and vectorization
$W \leftarrow []$                                          ▷ Set of time windows
$Y \leftarrow []$                                          ▷ Ground truth labels for windows
for $start\_time \leftarrow 0$ to $38,400$ step 600 do      ▷ Step 10 minutes
    $window\_states \leftarrow traffic\_states[start\_time : start\_time + 1,800]$
    $V_{concat} \leftarrow$ Concatenate$(window\_states)$
    $V_{avg} \leftarrow$ Average$(window\_states, axis = time)$
    $V_{concat} \leftarrow$ ZScoreNormalize$(V_{concat})$
    $V_{avg} \leftarrow$ ZScoreNormalize$(V_{avg})$
    $W.$append$((V_{concat}, V_{avg}))$
    $y \leftarrow$ DetermineGroundTruthLabel$(schedule, start\_time, start\_time + 1,800)$
    $Y.$append$(y)$
// STAGE 4: Dataset formation
$D \leftarrow \{(W[i], Y[i]) \mid i = 1, \ldots, 66\}$
// STAGE 5: Applying clustering algorithms
$Results \leftarrow \{\}$
for each algorithm $A \in \mathcal{A}$ do
    for each $data\_type \in \{concatenated, averaged\}$ do
        Perform Clustering and Calculate Metrics (ARI, Silhouette, etc.)
        $Results[(A, data\_type)] \leftarrow \{dataset : C_A, metrics : metrics\}$
return $D, Results$
```

termination. BayesianGMM is configured with $n\_components = 20$ and full covariance type. AgglomerativeClustering uses distance threshold 0.15. HDBSCAN is applied with cosine metric and $min\_cluster\_size = 4$. K-Means is tested with 5 and 7 clusters.

3.3.6. Creating ground truth labels

Ground truth labels are formed based on the simulation time schedule, where each of 66 time windows receives the label of the corresponding traffic scenario according to its activity period. Scenario time boundaries are determined considering window overlap and the need to ensure sufficient observation count for each pattern type.

3.4. Evaluation metrics

Clustering quality was evaluated by two metric categories. Internal metrics (Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index) characterize geometric properties of formed clusters without using external information. External metrics (V-measure, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Fowlkes-Mallows Score) compare clustering results with ground truth labels created by experts.

3.5. Experimental setup

Experiments were conducted on simulation data generated in SUMO 1.15.0 using a real city map. Clustering was performed using scikit-learn 1.3.0 in Python 3.9 environment. Ground truth labels were created based on simulation time schedule, where each of 66 windows received the label of corresponding scenario according to activity period.

4. Results

4.1. Experimental results presentation

4.1.1. Internal Clustering Quality Metrics

Table 1 demonstrates algorithm evaluation results by internal metrics, which do not require ground truth labels and characterize geometric properties of formed clusters.

Table 1
Internal clustering quality metrics for traffic trajectories.

Algorithm	Data Type	Silhouette	Davies-Bouldin	Calinski-Harabasz
AffinityPropagation	Averaged	0.54	0.84	277.91
AffinityPropagation	Concatenated	0.21	1.80	18.74
MeanShift	Averaged	0.44	0.64	151.52
MeanShift	Concatenated	0.16	1.31	14.11
BayesianGMM	Averaged	0.42	0.89	213.31
BayesianGMM	Concatenated	0.17	1.60	13.14
AgglomerativeClustering	Averaged	0.44	0.81	115.53
AgglomerativeClustering	Concatenated	0.05	0.53	3.59
HDBSCAN	Averaged	0.52	0.92	124.95
HDBSCAN	Concatenated	0.26	1.49	42.83
K-Means (5)	Averaged	0.54	0.62	244.04
K-Means (5)	Concatenated	0.21	1.94	35.54
K-Means (7)	Averaged	0.56	0.79	279.58
K-Means (7)	Concatenated	0.20	1.88	26.58

4.1.2. External clustering quality metrics

Table 2 shows comparison results with ground truth labels of expert-verified scenarios, allowing evaluation of accuracy in recovering true traffic pattern structure. Figure 3 presents comparison of Silhouette Score and Adjusted Rand Index indicators for all algorithms. The graph demonstrates the advantage of averaged data (circles) over concatenated (triangles), and positioning of HDBSCAN and K-Means in the upper right part indicates their optimal balance between geometric cluster quality and accuracy of ground truth scenario recovery.

Figure 4 presents a heatmap of normalized values of five quality metrics for averaged data. HDBSCAN demonstrates the most balanced high indicators across all external metrics, while BayesianGMM shows critically low values across practically all criteria.

Table 2
External clustering quality metrics for traffic trajectories.

Algorithm	Data Type	V-measure	ARI	NMI	Fowlkes-Mallows
AffinityPropagation	Averaged	0.69	0.67	0.69	0.72
AffinityPropagation	Concatenated	0.59	0.46	0.59	0.54
MeanShift	Averaged	0.72	0.63	0.72	0.69
MeanShift	Concatenated	0.61	0.41	0.61	0.57
BayesianGMM	Averaged	0.56	0.35	0.56	0.51
BayesianGMM	Concatenated	0.56	0.33	0.56	0.46
AgglomerativeClustering	Averaged	0.62	0.56	0.62	0.65
AgglomerativeClustering	Concatenated	0.45	0.03	0.45	0.16
HDBSCAN	Averaged	0.79	0.73	0.79	0.78
HDBSCAN	Concatenated	0.64	0.61	0.64	0.68
K-Means (5)	Averaged	0.73	0.70	0.73	0.76
K-Means (5)	Concatenated	0.67	0.62	0.67	0.71
K-Means (7)	Averaged	0.70	0.63	0.70	0.70
K-Means (7)	Concatenated	0.66	0.59	0.66	0.67

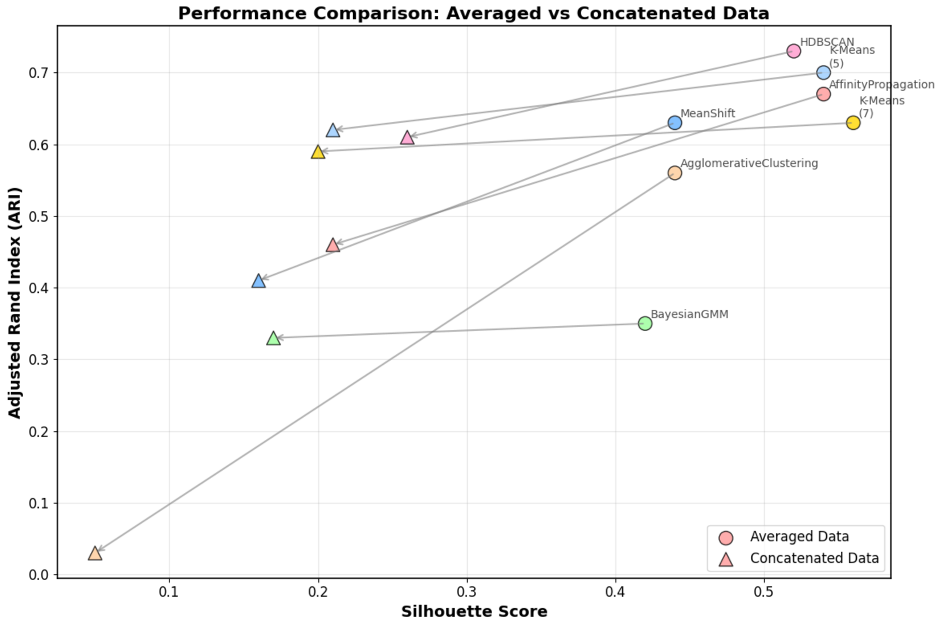

Figure 3: Performance comparison: Averaged data (circles) vs Concatenated data (triangles). Positioning in the upper right part of the graph indicates better clustering quality.

4.2. Results analysis

4.2.1. Algorithm comparison by internal metrics

Internal metrics analysis revealed a clear pattern: all algorithms demonstrate better results on aggregated (averaged) data compared to detailed (concatenated) values. K-Means with 7 clusters showed the highest Silhouette Score (0.56) and Calinski-Harabasz Index (279.58) for averaged values, indicating best geometric cluster quality. MeanShift demonstrated the lowest Davies-Bouldin Index (0.64) on averaged data, indicating optimal ratio of intra-cluster compactness and inter-cluster separation. Critical quality deterioration is observed for concatenated data: average Silhouette Score decrease is 0.25 points, and Calinski-Harabasz Index decreases on average by 8.7 times. AgglomerativeClustering showed the most

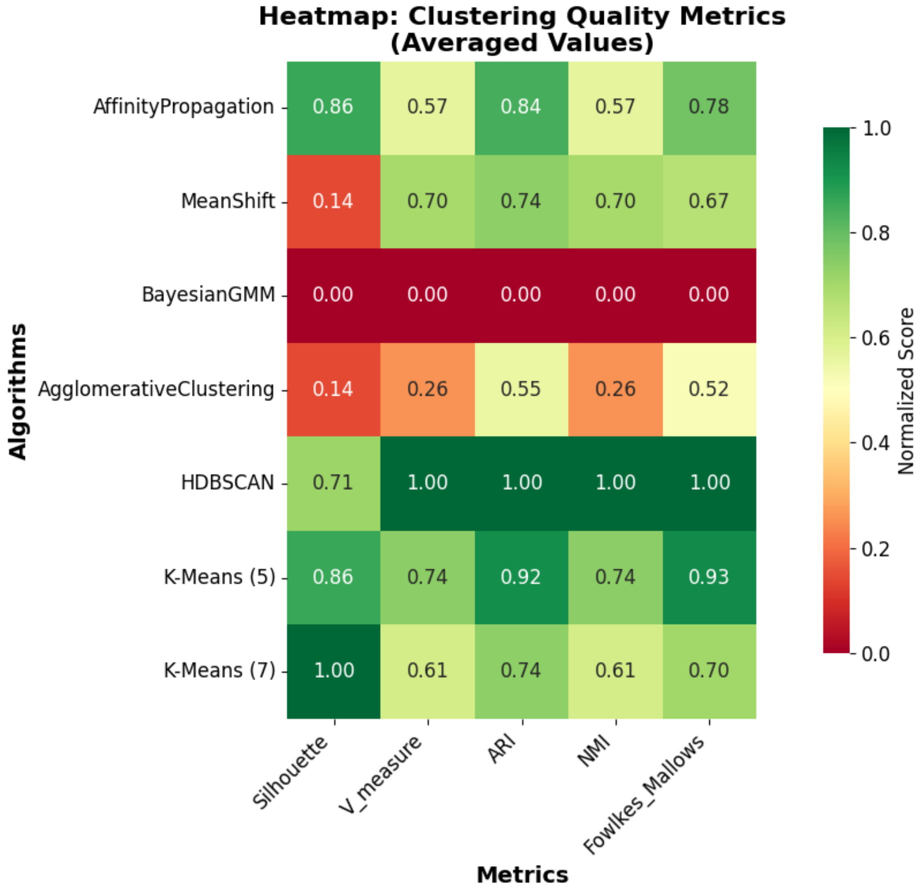

Figure 4: Heatmap of clustering quality metrics (averaged values). Green color intensity corresponds to higher quality; dark red color indicates low algorithm quality for the corresponding metric.

dramatic quality drop on concatenated data.

4.2.2. Ground truth scenario recovery accuracy

External metrics demonstrate HDBSCAN's advantage for accurate recovery of expert-verified traffic scenarios. HDBSCAN achieved the highest ARI (0.73) and V-measure (0.79) on averaged data, meaning 73% consistency with ground truth labels and balance between completeness and cluster homogeneity. K-Means (5 clusters) showed second place in accuracy (ARI = 0.70). MeanShift, despite high internal metrics, showed somewhat lower ground truth scenario recovery accuracy (ARI = 0.63). The worst results were demonstrated by AgglomerativeClustering on concatenated data (ARI = 0.03), practically corresponding to random point distribution across clusters.

5. Discussion

Results demonstrate clear advantage of aggregated (averaged) data over detailed (concatenated) values for all studied algorithms. HDBSCAN showed highest results by external metrics (ARI = 0.73, V-measure = 0.79), confirming its quality for traffic pattern identification. K-Means with 7 clusters achieved the highest Silhouette Score (0.56) but showed lower results in ground truth scenario recovery accuracy.

Significant quality deterioration on concatenated data (for example, HDBSCAN ARI decreases from 0.73 to 0.61) indicates that high dimensionality and temporal detail complicate stable pattern detection. AgglomerativeClustering showed critically low ARI (0.03) on concatenated data due to creating excessive numbers of small clusters.

Unlike existing research focusing on GPS trajectory analysis, our approach uses aggregated data from traffic light intersections, which better corresponds to practical urban traffic management needs. Results align with previous work conclusions regarding HDBSCAN advantages for traffic data but first demonstrate quantitative comparison on a controlled dataset.

Main limitations include: (1) using simulation data that may not fully reflect real traffic complexity; (2) limited number of scenarios (4 types) that may not cover all urban traffic pattern diversity; (3) focus on one city, limiting result generalizability; (4) absence of considering external factors (weather, events, accidents).

6. Conclusion

The research demonstrated HDBSCAN's capability for traffic pattern identification based on expert-verified simulation data, achieving the highest ground truth scenario recovery accuracy (ARI = 0.73, V-measure = 0.79) on aggregated data. Key numerical results show a significant advantage of using averaged values over detailed time series, with an average ARI improvement of 0.15 for all algorithms. HDBSCAN outperformed the baseline K-Means by 0.03 in ARI and 0.06 in V-measure, proving the effectiveness of density-based clustering in this domain. A critical finding is the susceptibility of concatenated, high-dimensional data to the curse of dimensionality, which drastically reduced the performance of algorithms like AgglomerativeClustering. The main limitation of this study lies in using simulation data from a single city with a limited set of four scenarios, which may restrict the generalizability of results to other urbanized territories with different topological complexity. To address this, future research expansion is planned through integrating real camera data to validate simulation findings, testing the methodology on multiple cities to ensure robustness, and developing hybrid approaches to improve mixed traffic scenario identification. These steps will further refine the benchmark methodology, enabling more effective traffic management systems capable of significant CO_2 emission reductions.

Funding

This research was funded by the European Union's Horizon Europe Framework Programme under grant agreement No. 101148374, project "U_CAN: Ukraine towards Carbon Neutrality." The views and opinions expressed are the authors' own and do not necessarily reflect those of the European Union or the funding agency, the European Climate, Infrastructure and Environment Executive Agency.

Acknowledgments

The authors would like to express their gratitude to the European Union's Horizon Europe Framework Programme for the financial support that made this research possible. Wealso extend our sincere appreciation to the developers and open-source communities behind the essential software tools used in this study, including SUMO, scikit-learn, pandas, and NumPy, whose contributions were invaluable to our work.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] International Energy Agency, Transport - energy system, 2024. URL: https://www.iea.org/energy-system/transport.

[2] N. Rouky, A. Bousouf, O. Benmoussa, M. Fri, A spatiotemporal analysis of traffic congestion patterns using clustering algorithms: A case study of casablanca, Decision Analytics Journal 10 (2024) 100404. doi:10.1016/j.dajour.2024.100404.

[3] I. T. Sarteshnizi, M. Sarvi, S. A. Bagloee, N. Nassir, Temporal pattern mining of urban traffic volume data: a pairwise hybrid clustering method, Transportmetrica B: Transport Dynamics (2023). doi:10.1080/21680566.2023.2185496.

[4] H. Ulvi, M. A. Yerlikaya, K. Yildiz, Urban traffic mobility optimization model: A novel mathematical approach for predictive urban traffic analysis, Applied Sciences 14 (2024) 5873. doi:10.3390/app14135873.

[5] I. Portugal, P. Alencar, D. Cowan, A framework for spatial-temporal cluster evolution representation and analysis based on graphs, Scientific Reports 14 (2024) 5873. doi:10.1038/s41598-024-72504-x.

[6] I. Krak, O. Barmak, E. Manziuk, Using visual analytics to develop human and machine-centric models: A review of approaches and proposed information technology, Computational Intelligence 38 (2022) 921–946. doi:10.1111/coin.12289.

[7] D. Wang, Y. Huang, Z. Cai, A two-phase clustering approach for traffic accident black spots identification: integrated gis-based processing and hdbscan model, International Journal of Injury Control and Safety Promotion (2023). doi:10.1080/17457300.2022.2164309.

[8] T. Yang, L. Wang, L. Zhou, H. Chen, A detection of multi-level co-location patterns based on column calculation and hdbscan clustering, Intelligent Data Analysis (2025). doi:10.1177/1088467X241308765.

[9] D. M. Bot, J. Peeters, J. Liesenborgs, J. Aerts, Flasc: a flare-sensitive clustering algorithm, PeerJ Computer Science 11 (2025) e2792. doi:10.7717/peerj-cs.2792.

[10] G. Monko, M. Kimura, Enhanced stratified sampling-density-based spatial clustering of applications with noise (ss-dbscan) for high-dimensional data, Data Science 8 (2025). doi:10.1177/24518492251349080.

[11] F. Gonçalves, G. O. Silva, A. Santos, A. M. A. C. Rocha, H. Peixoto, D. Durães, J. Machado, Urban traffic simulation using mobility patterns synthesized from real sensors, Electronics 12 (2023) 4971. doi:10.3390/electronics12244971.

[12] C. Stang, K. Bogenberger, Calibration of microscopic traffic simulation in an urban environment using gps-data, in: SUMO Conference Proceedings, volume 5, 2024, pp. 71–78. doi:10.52825/scp.v5i.1099.

[13] A. Keler, A. Kunz, S. Amini, K. Bogenberger, Calibration of a microscopic traffic simulation in an urban scenario using loop detector data: A case study within the digital twin munich, in: SUMO Conference Proceedings, volume 4, 2023, p. 153. doi:10.52825/scp.v4i.223.

[14] X. Wang, X. Dong, Z. Zhang, Y. Wang, Transportation carbon reduction technologies: A review of fundamentals, application, and performance, Journal of Traffic and Transportation Engineering (English Edition) 11 (2024) 1340–1377. doi:10.1016/j.jtte.2024.11.001.

[15] I. Derpich, C. Duran, R. Carrasco, F. Moreno, C. Fernandez-Campusano, L. Espinosa-Leal, Pursuing optimization using multimodal transportation system: A strategic approach to minimizing costs and co2 emissions, Journal of Marine Science and Engineering 12 (2024). doi:10.3390/jmse12060976.

[16] G. M. I. Alam, S. A. Tanim, S. K. Sarker, Y. Watanobe, R. Islam, M. F. Mridha, K. Nur, Deep learning model based prediction of vehicle co2 emissions with explainable ai integration for sustainable environment, Scientific Reports 15 (2025). doi:10.1038/s41598-025-87233-y.

[17] M. Mądziel, Predictive methods for co2 emissions and energy use in vehicles at intersections, Scientific Reports 15 (2025) 6463. doi:10.1038/s41598-025-91300-9.

[18] K. Wu, J. Ding, J. Lin, G. Zheng, Y. Sun, J. Fang, T. Xu, Y. Zhu, B. Gu, Big-data empowered traffic signal control could reduce urban carbon emission, Nature Communications 16 (2025) 2013. doi:10.1038/s41467-025-56701-4.

[19] O. Ryzhanskyi, V. Pavlyshyn, P. Radiuk, E. Manziuk, O. Barmak, I. Krak, Ai-driven traffic signal control system to reduce co2 emissions, in: CEUR Workshop Proceedings, volume 3974, 2025, pp.

18–27.

[20] V. Pavlyshyn, E. Manziuk, O. Barmak, I. Krak, R. Damasevicius, Modeling environment intelligent transport system for eco-friendly urban mobility, in: CEUR Workshop Proceedings, volume 3675, 2024, pp. 118–136. URL: https://ceur-ws.org/Vol-3675/paper9.pdf.

[21] Y. Bie, Y. Ji, D. Ma, Multi-agent deep reinforcement learning collaborative traffic signal control method considering intersection heterogeneity, Transportation Research Part C: Emerging Technologies 164 (2024) 104663. doi:10.1016/j.trc.2024.104663.

[22] D. Li, F. Zhu, J. Wu, Y. D. Wong, T. Chen, Managing mixed traffic at signalized intersections: An adaptive signal control and cav coordination system based on deep reinforcement learning, Expert Systems with Applications 238 (2024). doi:10.1016/j.eswa.2023.121959.

[23] P. Radiuk, N. Hrypynska, A framework for exploring and modelling neural architecture search methods, in: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (CoLInS 2020), volume 2604, CEUR-WS.org, Aachen, 2020, pp. 1060–1074. URL: https://ceur-ws.org/Vol-2604/paper70.pdf.

[24] S. Afandizadeh, S. Abdolahi, H. Mirzahossein, Deep learning algorithms for traffic forecasting: A comprehensive review and comparison with classical ones, Journal of Advanced Transportation (2024). doi:10.1155/2024/9981657.

[25] M. Berlotti, S. Di Grande, S. Cavalieri, Proposal of a machine learning approach for traffic flow prediction, Sensors 24 (2024) 2348. doi:10.3390/s24072348.

[26] S. Mishra, T. S. Murthy, A predictive and optimization approach for enhanced urban mobility using spatiotemporal data, arXiv preprint arXiv:2410.05358 (2024). doi:10.48550/arXiv.2410.05358.

[27] Z. Huang, W. Wang, S. Huang, M. C. Gonzalez, Y. Jin, Y. Xu, Where to go next day: Multi-scale spatial-temporal decoupled model for mid-term human mobility prediction, arXiv preprint arXiv:2501.06561 (2025). doi:10.48550/arXiv.2501.06561.

[28] J. Liu, Y. Yuan, Exploring dynamic urban mobility patterns from traffic flow data using community detection, Annals of GIS 30 (2024) 435–454. doi:10.1080/19475683.2024.2324393.

[29] O. Yusuf, A. Rasheed, F. Lindseth, Exploring urban mobility trends using cellular network data, in: The 1st International Conference on Net-Zero Built Environment, 2025, pp. 1661–1674. doi:10.1007/978-3-031-69626-8_138.

[30] Y. Wang, F. Currim, S. Ram, Deep learning of spatiotemporal patterns for urban mobility prediction using big data, Information Systems Research 33 (2022) 579–598. doi:10.1287/isre.2021.1072.

[31] H. Amdouni, G. Manita, D. Oliva, E. H. Houssein, O. Korbaa, S. Zapotecas-Martínez, Dynamic social particle swarm optimization for automatic clustering, Procedia Computer Science 246 (2024) 1409–1418. doi:10.1016/j.procs.2024.09.583.

[32] A. M. Ikotun, F. Habyarimana, A. E. Ezugwu, Benchmarking validity indices for evolutionary k-means clustering performance, Scientific Reports 15 (2025) 21842. doi:10.1038/s41598-025-08473-6.

[33] O. Barmak, I. Krak, E. Manziuk, Diversity as the basis for effective clustering-based classification, in: CEUR Workshop Proceedings, volume 2711, 2020, pp. 53–67. URL: http://ceur-ws.org/Vol-2711/paper5.pdf.

[34] O. Barmak, Y. Krak, E. Manziuk, Characteristics for choice of models in the ensembles classification, in: CEUR Workshop Proceedings, volume 2139, 2018, pp. 171–179. URL: https://ceur-ws.org/Vol-2139/171-179.pdf.