Methodology for the Identification of Vehicle Congestion Based on Dynamic Clustering

Gary Reyes *, Roberto Tolozano-Benites, Laura Lanzarini, César Estrebou, Aurelio F. Bariviera, Julio Barzola-Monteses

Posted Date: 16 November 2023

doi: 10.20944/preprints202311.1043.v1

Keywords: congestion; dynamic clustering; GPS trajectories; road networks

Article

Methodology for the Identification of Vehicle Congestion Based on Dynamic Clustering

Gary Reyes [1,2,*,†], Roberto Tolozano-Benites [1,†], Laura Lanzarini [3,†], César Estrebou [3,†], Aurelio Bariviera [4,†] and Julio Barzola-Monteses [1,2,†]

1 Universidad Bolivariana del Ecuador, Campus Durán Km 5.5 vía Durán Yaguachi, 092405 Durán, Ecuador
2 Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Cdla. Universitaria Salvador Allende, Guayaquil 090514, Ecuador
3 Universidad Nacional de La Plata, Facultad de Informática,Instituto de Investigación en Informática LIDI (Centro CICPBA) 1900 La Plata, Buenos Aires, Argentina
4 Universitat Rovira i Virgili, Department of Business, Reus, Spain
* Correspondence: gxreyesz@ube.edu.ec
† These authors contributed equally to this work.

Abstract: Addressing sustainable mobility in urban areas has become a priority in today's society, given the growing population and increasing vehicular flow in these areas. Intelligent Transportation Systems have emerged as innovative and effective technological solutions to address these challenges. Research in this area has become crucial, as it contributes not only to improve mobility in urban areas, but also to positively impact the quality of life of its inhabitants. To address this, a dynamic clustering methodology for vehicular trajectory data is proposed which can provide an accurate representation of the traffic state. Data was collected for the city of San Francisco, a dynamic clustering algorithm was applied and then an indicator was applied to identify areas with traffic congestion. Several experiments were also conducted with different parameterizations of the forgetting factor of the clustering algorithm. The results showed in terms of precision that the dynamic clustering methodology achieved high match rates compared to the congestion indicator applied to static cells.

Keywords: congestion; dynamic clustering; GPS trajectories, road networks

1. Introduction

Sustainable mobility has emerged in response to the environmental and social challenges associated with urban growth and increased vehicular traffic. This paradigm seeks to transform modes of travel, promoting alternatives that reduce greenhouse gas emissions and minimize the impact on ecosystems. Research efforts have focused on a variety of fronts, from the development of more efficient and cleaner vehicle technologies to mobility-oriented urban planning. Sustainable mobility research has become an interdisciplinary and constantly evolving field, driven by the urgent need to find viable and sustainable solutions to the growing transportation demands in cities.

The most crucial research areas, revealing a diverse and complex landscape, have emerged as a crucial convergence with Intelligent Transportation Systems (ITS), marking a transition towards more advanced and effective solutions. These systems, supported by innovative technologies such as real-time data analytics and artificial intelligence, offer unprecedented opportunities to improve traffic management, facilitate urban planning, and encourage the adoption of sustainable modes of transportation.

Research and analysis of Intelligent Transportation Systems in urban areas have become essential today due to the complexity of this problem and its profound impact on society. The constant growth of the population in urban areas, as well as the increase in vehicular traffic, are obvious factors that require careful attention [1].

Intelligent Transportation Systems emerge as an innovative and technological response to address these challenges in an efficient and sustainable manner. In their search for effective solutions to the

 doi:10.20944/preprints202311.1043.v1

challenges of urban mobility, they employ a variety of machine learning techniques to obtain practical applications and offer analytical approaches in the field of transportation.

Intelligent Transportation Systems use machine learning algorithms to detect patterns in vehicle behavior, such as regular congestion in certain areas or drivers' preferred routes. This information is essential for congestion prediction, optimal route planning and real-time adaptation of traffic management strategies.

The applicability of these approaches is broad, ranging from real-time traffic management to long-term planning of transportation infrastructure. By better understanding traffic patterns and driver behaviors, intelligent systems can offer more effective solutions, such as traffic light optimization, public transport route management, and the implementation of sustainable mobility policies.

In this regard, the management of vehicular traffic in urban areas is of great importance due to the constant population growth and increase in vehicles, which poses significant challenges [2]. This management must address multiple dimensions, including environmental impact and road safety. Traffic congestion is a recurring problem that affects the quality of life of citizens.

There are challenges in managing traffic congestion such as the lack of an accurate and uniform representation of vehicle trajectory data that makes early identification of congested areas difficult [3]. Dispersion and incompleteness of data collection points are also common problems.

Efficient traffic management is essential to improve road flow, reduce travel time and reduce pollutant emissions. Traditional approaches may not adapt quickly to changing traffic conditions, which is essential given that congestion can vary significantly at different times of the day [4].

Data streams, collected from various sources such as traffic sensors and GPS navigation systems, are essential for understanding the real-time behavior of vehicles and pedestrians in urban areas [5–7]. Clustering techniques are valuable for representing these data streams effectively, allowing identification of traffic patterns, organization of data into clusters based on similarities, and prediction of future trends in urban traffic. These techniques are fundamental for traffic planning and management tailored to the specific needs of each area.

The analysis of vehicular trajectory data streams is a widely researched area [8], and several studies have developed clustering techniques adapted to different domains [9–11]. The study of various approaches has proven effective in identifying sets with shared attributes in the analysis of the joint behavior of vehicles [12,13].

Some researchers have adapted conventional clustering methods, such as k-means [9] and DBSCAN [14], by adapting methods and calculations designed specifically for trajectories [15]. Several investigations have resorted to alternative representation [16] of trajectories such as subdvision or cell representation to improve clustering results [17,18].

In some cases static vehicle analysis may be limited in its ability to capture real traffic dynamics. Because vehicle behavior can change over time [19], dynamic analysis has become important for understanding the causes of congestion [20]. In recent years, there has been an increase in artificial intelligence and machine learning approaches that add features such as memory, scalability and accuracy [21–23]. Machine learning has proven its effectiveness by leveraging the use of historical information combined with information associated with vehicles and the road environment in which they travel [24–26]. These combinations, enriched by the inclusion of data from Big Data, especially generated from social networks, have become an invaluable resource for detecting traffic congestion in real time [27].

Several studies have developed methodologies and techniques to identify congested areas accurately, using a variety of traffic and environmental characteristics [28–33].

Several proposals with combined approaches focusing on traffic congestion assessment and the use of clustering algorithms constitute a highly promising field of research [34–36], providing an effective method to closely examine vehicular flow in different scenarios [28,37,38].

The paper proposed by Almeida et al. [39] proposes a method for traffic congestion detection considering speed, traffic flow and road occupancy and then uses clustering techniques to detect

 doi:10.20944/preprints202311.1043.v1

various degrees of congestion in vehicular data. The paper proposed by Reyes et al. [40] analyzes vehicular flow by identifying speed ranges with a constant update. Although it is a simplified view of vehicular traffic flow, in many cases it is beneficial to include additional data in order to enrich the study of vehicular traffic [41].

A detailed understanding of how congestion manifests and evolves in different environments is critical to strategically plan mobility, alleviate congestion, and ensure more efficient and sustainable traffic flow.

This paper proposes a methodology to analyze vehicular flow by clustering vehicle trajectory data with GPS points. This methodology allows an accurate representation of the data, especially useful when points are scarce. It uses clusters to detect areas of congestion patterns. The constant updating of the clusters ensures up-to-date data and real congestion management. It also uses a congestion indicator to measure traffic saturation allowing a dynamic view of the traffic situation in different areas.

This article is organized as follows: Section 2 describes the proposed methodology, Section 3 present the obtained results, Section 4 discusses the obtained results, and Section 5 presents the conclusions and future lines of work.

2. Materials and Methods

This paper presents a methodology for the identification of congestion zones based mainly on dynamic clustering. The methodology used consists of five steps illustrated in Figure 1. In the first step, road network data information is loaded. In the second step, a GPS trajectory data stream is processed. In the third step, a distance-based dynamic clustering algorithm is used to identify areas with similar patterns. In the fourth step, areas are evaluated with a congestion indicator for further classification. In the fifth step, we proceed to generate a suitable visualization of the resulting clusters already classified. Each of these steps is described in detail below.

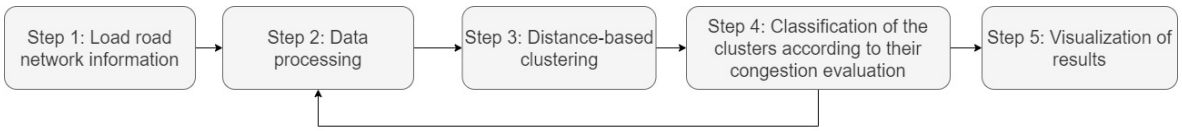


Figure 1. Steps of the proposed methodology.

2.1. Step 1: Load Road Network Information

The estimation of reference data is important to ensure the reliability of the results and to strengthen the validity and consistency of the analyses performed.

The main purpose of this step is to load into memory the relevant information on the road infrastructure in the area to be analyzed. This includes detailed data on road layout, geographical location, capacity, number of lanes and speed limits. These data will not only be used for the analysis, but will also enable an effective comparison between different road sections, thus establishing a solid benchmark for detecting and assessing road congestion accurately.

The robustness of the results depends on the quality and comprehensiveness of the data collected in this step. Accuracy in the representation of the road network and thoroughness in data collection are essential to ensure robust and reliable results in the subsequent steps of this study.

2.2. Step 2: Data Processing

In this step, a method is established to receive and process GPS points in real time or from an accessible repository. This data comes from GPS-equipped vehicles or in-vehicle mobile applications. Processing is performed in microbatches at evenly distributed time intervals, called "cycles", which represent moments in the evolution of the data streams.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2023 doi:10.20944/preprints202311.1043.v1

At each cycle, data is accumulated in a temporal buffer, and the results of the clustering method are updated as new data is added. To address the lack of information in the trajectory data, a routing method and an interpolation method are implemented that operate simultaneously during the collection of new GPS locations.

The decision to perform routing and interpolation along with data accumulation is justified by its ability to generate an enriched and densely populated data buffer. This ensures that as the data stream progresses, it works with complete and accurate route information for each vehicle, improving the efficiency of the subsequent clustering analysis.

The routing method uses the Open Source Routing Machine (OSRM) service, based on contraction hierarchies and travel time optimization. This service calculates the shortest route and generates key spatial points based on the geometry of the road networks, projecting an artificial route on the roads that the vehicle could have driven on.

The distance-based interpolation method consists of continuously estimating data values along a route with a constant interval of 5 meters. These techniques are applied to complete routes in cases where there are no records for certain sections, providing a more detailed and uniform representation of the data, especially useful when collection points are scattered or incomplete.

Routing and interpolation generate a denser and more continuous data set, allowing for accurate analysis and seamless visualization of vehicle behavior along the route. This interpolated information is valuable for identifying and analyzing trends for individual vehicles, as illustrated in Figure 2.

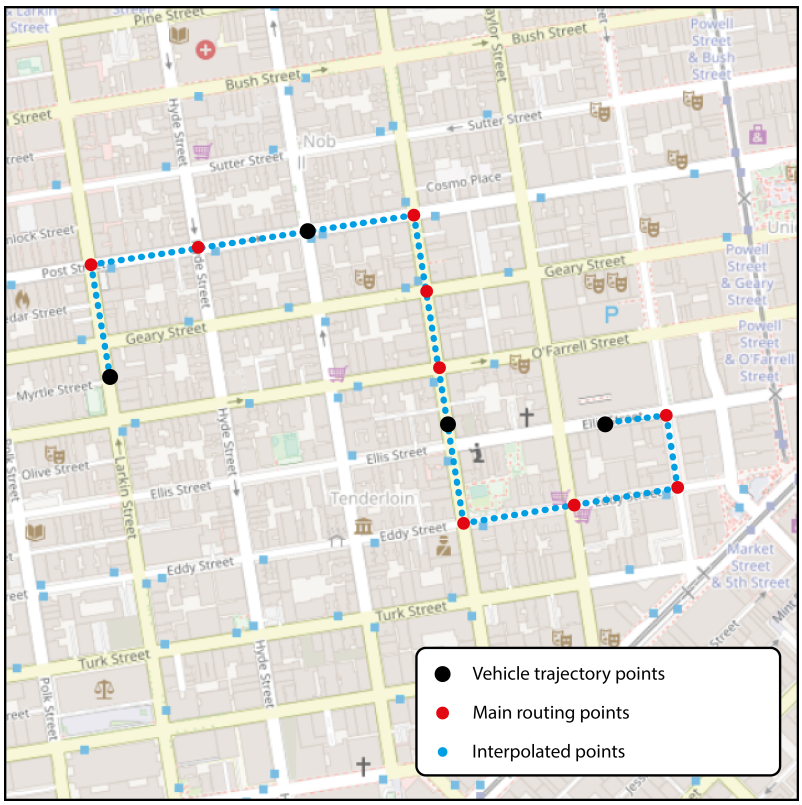


Figure 2. Routing and interpolation of a trajectory.

2.3. Step 3: Distance-based Clustering

A cluster is mainly composed of a centroid, a hyperbox and linked GPS points. The centroid, which is the geographic point representing the center of the cluster, is used as a representative reference for the cluster in the analysis. The hyperbox is a rectangular structure, it is positioned according to the cluster centroid. This rectangular shape delimits an area around the centroid, which simplifies the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2023 doi:10.20944/preprints202311.1043.v1

spatial representation and determines its area of influence. The visual representation of a cluster is presented in Figure 3.

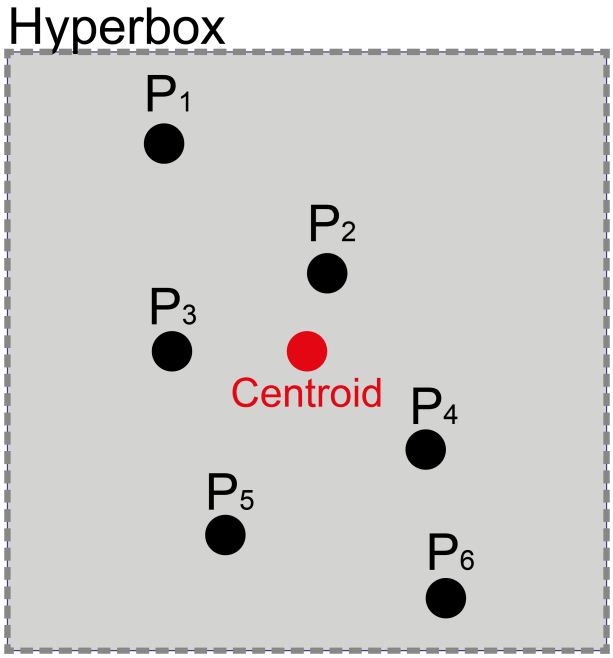


Figure 3. Elements that make up a cluster.

Each GPS point processes geographic location information, vehicle identification and time of entry. Clustering is performed using similarity based on Euclidean distance, considering the latitude and longitude attributes of the GPS points. Each GPS point is analyzed by calculating the Euclidean distance with the centroids of the existing clusters. Each point is assigned to the cluster with the smallest spatial distance and within the hyperbox area.

In case it is not in the hyperbox area, a new cluster is created. Points assigned to a cluster cannot be reassigned to another cluster. The centroid is updated when new GPS points are integrated into a cluster, and new clusters are created if there are no nearby clusters.

To ensure that the clusters are updated and to avoid retention of old data, a forgetting method based on the time of entry of the last GPS point is used to determine the loss of relevance as time elapses and is calculated by Equation 1.

$$F = e^{-1*\lambda*\delta t} \tag{1}$$

where, e represents the exponential function, λ controls the speed of the loss in relevance, and δt is the difference between the time of the analyzed point and the time of the last point integrated to the cluster.

A relevance threshold of 5% has been established to determine when a value is no longer relevant to the analysis. This method determines the number of GPS points that will remain in the cluster during the retention period, thus adapting the cluster to changes in traffic and avoiding the accumulation of obsolete data. Clusters that lose relevance due to lack of new GPS points are deleted, while active clusters that continue to receive data are kept up to date.

2.4. Step 4: Classification of the Clusters according to their Congestion Assessment

Once all the points in the temporary buffer have been assigned to a cluster, each cluster formed is evaluated by means of an indicator that analyzes the state of the traffic based on the results of the clustering at that moment. The temporary buffer can be further used to process a new data flow with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2023 doi:10.20944/preprints202311.1043.v1

its respective clustering. This step allows to analyze and classify each cluster individually according to its congestion level, which will help to identify problem areas and areas with better traffic flow.

Each cluster is examined individually to understand its behavior and particular characteristics, these characteristics are reflected through statistical measurements which are implicit in each road segment found within the cluster area, have been present since its initial formation and are updated each time the cluster presents changes, from the segments information can be obtained on the number of GPS points, the average speed of vehicles (unit measured in kilometers per hour, km/h), the number of vehicles, among others.

A hyperbox is spatially projected for each cluster to delimit its respective area of analysis. Map data are used to identify the roads and road segments contained within the delimited area of each cluster. A spatial clipping is applied to the roads to fit within the area defined by the hyperbox for each cluster, this allows isolating the relevant road sections that influence each specific cluster, an example of this clipping can be seen in Figura 4. Additionally, information associated with the geometry and metadata of each road will be used.

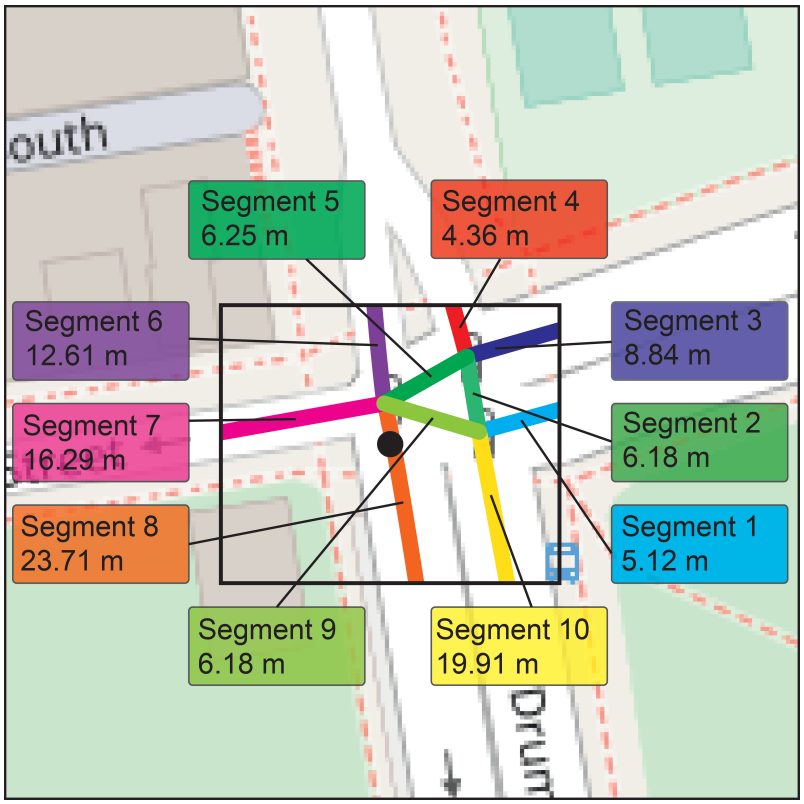


Figure 4. Road segments of the analyzed area of a cluster.

For the classification of the congestion status of the clusters, a Traffic Coefficient Indicator has been used as a congestion indicator. The congestion indicator used indicates a value that reflects the level of congestion at a location or road [42], quantitatively measuring congestion based on the density of vehicles and their speeds. A high value of this congestion indicator indicates significant congestion, while a low value of the congestion indicator suggests smooth traffic.

The calculation of this congestion indicator is done by road segments. Since a hyperbox of a cluster may comprise several segments, the congestion indicator is calculated individually for each segment. Then, a unified value is generated based on the length of the segments with at least one vehicle so that each cluster has its own value generated by the congestion indicator.

The value of the congestion indicator is calculated by the relationship between a Density Index and a Speed Index.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2023 doi:10.20944/preprints202311.1043.v1

The Density Index represents the number of vehicles on a road segment at a specific time. It is calculated by dividing the number of vehicles observed in the area by the maximum number previously recorded on its respective road segment.

This maximum amount is based on historical data or previously conducted traffic studies. In this research, a systematic procedure for the dynamic determination of the maximum traffic density value is established. The procedure begins by identifying the road segments in the study area, as shown in Figure 5, and calculating the traffic density of each. These densities are converted into density per unit length (D/L) values, considering the different lengths of the segments.

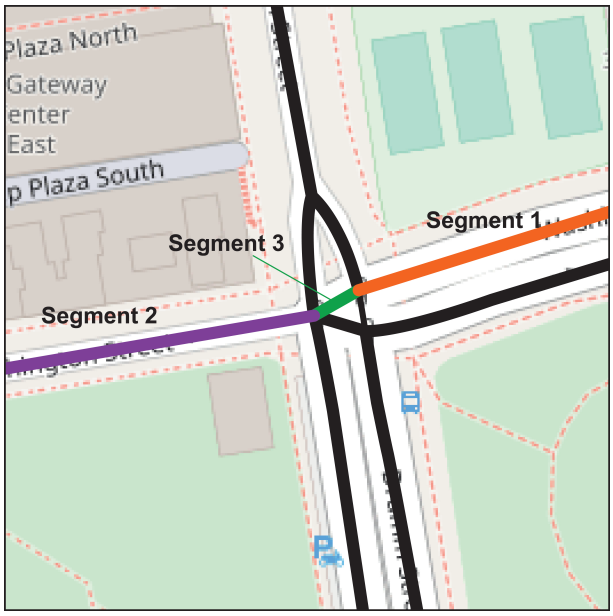


Figure 5. Segments identified from the road network.

Then, the proportion of each segment is calculated as a function of its length with respect to the total number of segments traveled. The weighted values of density per unit length are obtained by multiplying the D/L with the proportions of each segment and adding them together to obtain a representative measure of the cluster under consideration.

The densities of all clusters in the cycle are then used and averaged to obtain an overall value for that cycle. This value is added to a historical record that is updated each cycle. With this historical value, the maximum traffic density can be estimated in a generalized manner for different road segment lengths by multiplying the historical value by the length of the road under analysis. The inclusion of multiple road segment lengths in the historical record ensures accuracy and reliability in determining the maximum traffic density, regardless of the length of the segment under analysis.

When the Density Index approaches or reaches 1, it indicates that the number of vehicles in that area is close to or has exceeded the maximum observed capacity. This suggests a high probability of congestion.

The Speed Index reflects the average speed of vehicles on analyzed roads. It is calculated by dividing the average vehicle speed by the speed limit set by local traffic regulations. These regulations are determined according to the regulations of each city, with the purpose of ensuring adequate traffic flow.

It is calculated by dividing the average speed of vehicles observed on each segment by the maximum speed allowed on the respective road segment.

When the Speed Index is close to or equal to 1, vehicles are traveling at the maximum allowable speed, indicating smooth traffic flow and low congestion. Conversely, a lower speed indicates a slower flow of traffic, which could indicate the presence of congestion.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2023 doi:10.20944/preprints202311.1043.v1

2.5. Step 5: Results Visualization

In this study, trajectory information is examined at regular intervals, allowing accurate detection of changes in vehicular flow.

In order to provide a visual and interactive representation of the results of each cluster, an interactive map has been developed that can be generated in any cycle. This map allows dynamic and graphical analysis of the relevant information for each cluster. Each area with similar characteristics is represented with a different color on the map, as illustrated in Figure 6.

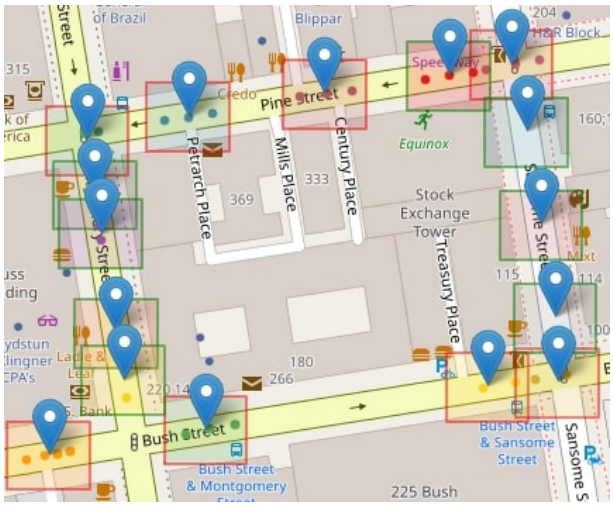


Figure 6. Clusters projected on the map.

3. Results

3.1. Used Data

3.1.1. San Francisco Dataset

The data for the city of San Francisco were obtained on June 02, 2008 and cover a total of 290 trajectories followed by cabs that were equipped with GPS positioning devices.

Each record includes detailed information such as a trajectory identifier, latitude and longitude coordinates, time information, speed and direction. In this set of trajectories, an analysis of all routes recorded during the time interval from 12:30 p.m. to 13:30 p.m. was conducted. After this selection process, 2382 records were obtained, representing all 290 trajectories contained in the original data set.

The area that represents the selected dataset is shown in Figure 7.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2023 doi:10.20944/preprints202311.1043.v1

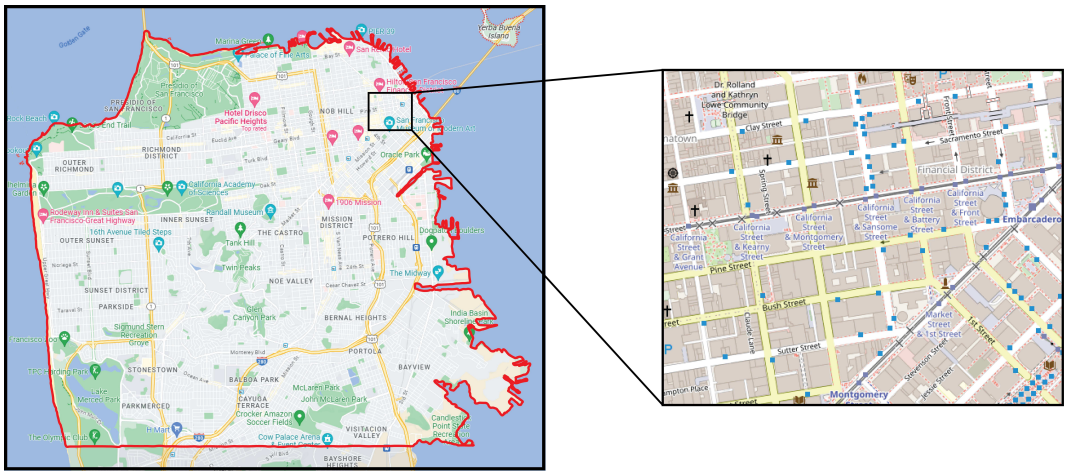


Figure 7. Area representing the dataset for the city of San Francisco.

3.2. Model Parameter Selection

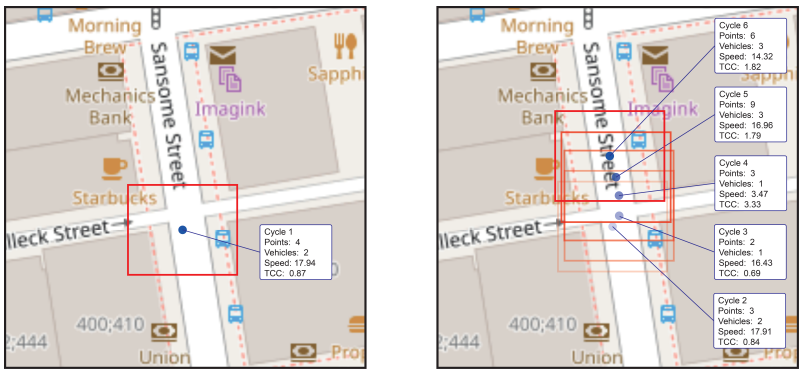
In this work, an analysis area of 1200x800 square meters has been delimited. The hyperboxes represent approximately 3% of the analysis area and have a size of about 35x25 meters. The analysis cycles have a duration of 1 minute each. Euclidean distance is used as the similarity measure, and the forgetting parameter was set to 0.068 and 0.05, to consider relevant data up to 45 and 60 seconds respectively. Clusters with low activity are updated every 30 seconds, and old clusters are removed if they have stopped receiving new points after 2 minutes.

3.3. Model Testing

To demonstrate the advantages of the dynamic clustering methodology compared to the congestion indicator applied to static cells, a model test was performed. The main objective of this test was to analyze how the methodology deals with the dynamics of traffic data and vehicular flow in a road network, identifying situations in which it is superior.

A data set representative of the city of San Francisco was used, comprising 6 run cycles in a 100x100 meter area. A cluster of data from this test was randomly selected for comparative analysis.

The dynamic clustering methodology excels in its ability to adapt to variations in data distribution. Figure 8a shows how the hyperbox was flexibly and accurately adjusted to encompass road segments, effectively capturing variations in density and shape of the clusters as the data evolved, as can be seen in Figure 8b. In addition, this methodology demonstrated a clear advantage in the selection of road segments subject to variations in vehicular flow and traffic density.



(a) (b)

Figure 8. Dynamism of the clusters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2023 doi:10.20944/preprints202311.1043.v1

In contrast, the congestion indicator applied to static cells uses a fixed grid., which is seen in Figure 9. This limits its adaptability and can compromise the quality of the results by not adjusting to changing patterns of vehicle trajectories. Not being able to adapt to how data is clustered over time makes it unsuitable for identifying congestion events in dynamic traffic scenarios.

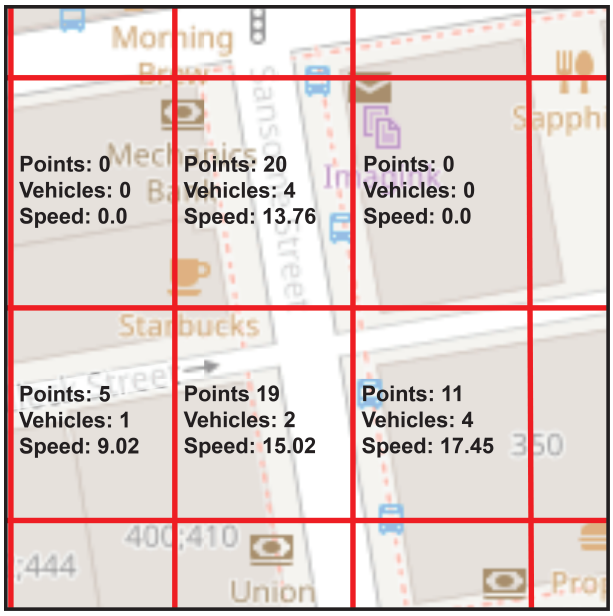
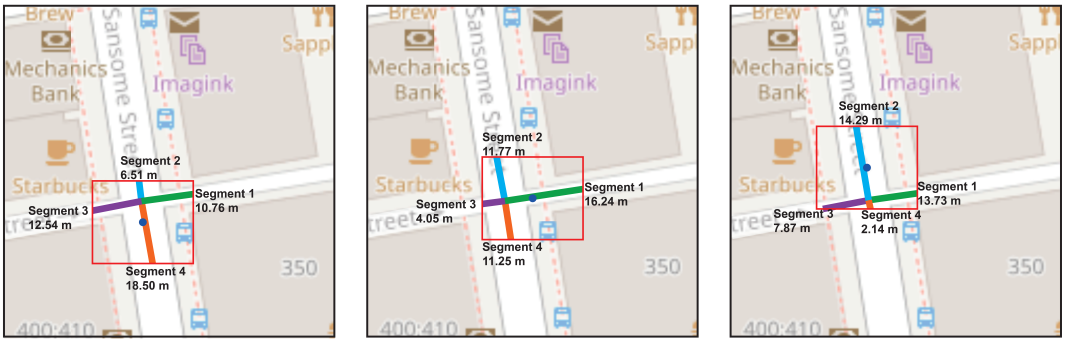


Figure 9. Representation of the cells in a static grid.

The dynamic clustering methodology also demonstrates a significant advantage in terms of adaptability to data dynamics. In Figure 10, its ability to select road segments in cycles 1, 3, and 5 is seen, allowing it to adjust to fluctuations and changes in the data distribution. This methodology can modify the position of the centroids and hyperbox as needed, making it a suitable choice for selecting road segments subject to variations in vehicular flow.



(a) (b) (c)

Figure 10. Dynamic road segment selection in cycle 1 (a), cycle 3 (b) and cycle 5 (c).

In contrast, the congestion indicator applied to static cells faces difficulties when dealing with these dynamics due to its fixed grid. In real scenarios, where roads experience fluctuations in the number of vehicles throughout the day, the congestion indicator applied to static cells may select incorrect road segments for each cell, as seen in Figure 11. This can lead to an incorrect representation of traffic at different times of the day.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2023 doi:10.20944/preprints202311.1043.v1

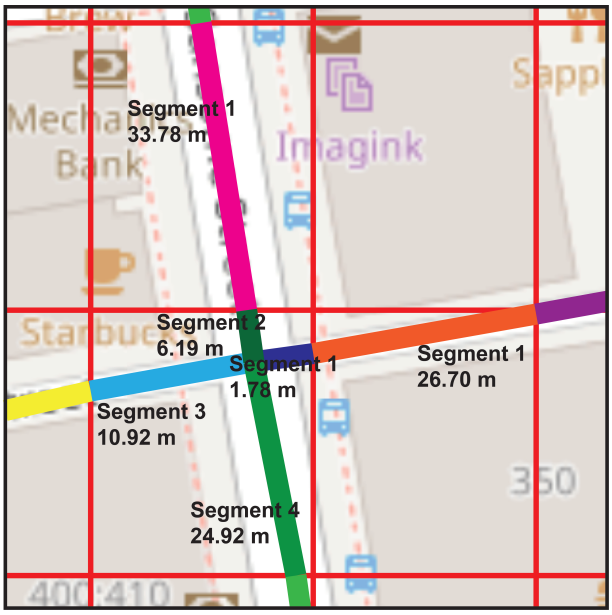


Figure 11. Selection of road segments for fixed cells.

A highlight of the dynamic clustering methodology is its ability to incorporate more recent vehicle locations in real time, resulting in an automatic update of the cluster centroids. This ensures that newer locations have a greater impact on defining real-time congestion, while older locations become less and less relevant. This is crucial, as a vehicle can cross multiple cells in a single trip, the dynamic clustering methodology ensures an accurate and sensitive assessment of congestion, adapting to the changing mobility of vehicles on urban roads.

3.4. Obtained Results

In this research, the results of the dynamic clustering methodology were compared with the results of the congestion indicator applied to static cells to analyze vehicular flow using the traffic coefficient indicator to measure congestion, as shown in Figure 12.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2023 doi:10.20944/preprints202311.1043.v1

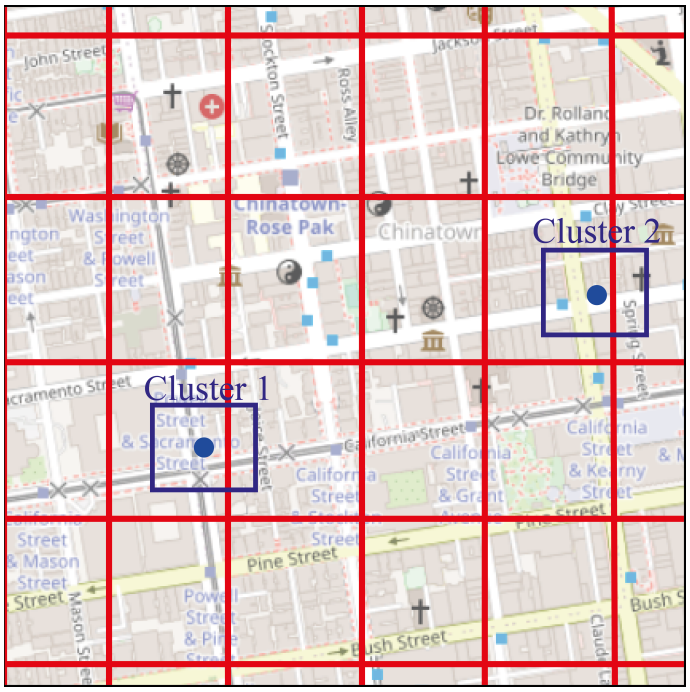


Figure 12. Cluster projected on the grid.

In the dynamic clustering methodology, the vehicle data were grouped into similar patterns and the location of the cluster determines the area of comparison, while in the static cells, the area was divided into uniform cells, both cases are evaluated using the congestion indicator.

A tolerance was applied to the indicator values in dynamic clusters to account for the natural variability of the data and to avoid misidentification of possible congestion states. The tolerance determines how much margin of error is allowed when adjusting the congestion indicator, directly influencing the number of matches observed in the congestion classification. The congested classification results were then compared between cells and clusters, recording valid matches when at least one cell matched the same congestion classification as the cluster.

To determine if the cluster classification has been performed correctly, the results of the model used using confusion matrices are compared to the results from the city of San Francisco. The following confusion matrices present a representation of the capabilities of the dynamic cluster classification model relative to the static grid model. The values reflect the number of accurate and erroneous predictions in the congested and non-congested categories. These results provide detailed insight into how the model interprets and predicts the congestion state in the clusters, which is essential for evaluating its performance in real traffic situations.

The confusion matrix for the city of San Francisco, with a forgetting value of 45 seconds and no tolerance, is presented in the Table 1, in the congestion cases, 6433 were correctly classified, and of the non-congested cases, 10932 were correctly classified. However, 1273 errors were made in misclassifying non-congested situations as congestion, and 6307 errors were made in classifying congested situations as non-congested.

Table 1. Confusion matrix for the city of San Francisco using 45 seconds of forgetting and no tolerance.

	Congested cells	**Non-congested cells**
Congested clusters	6433	1273
Non-congested clusters	6307	10932

The confusion matrix for the city of San Francisco, with a forgetting value of 45 seconds and a tolerance of 0.2, is presented in Table 2, in the case of congested clusters, 6881 cases were correctly

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2023 doi:10.20944/preprints202311.1043.v1

classified as congestion, and in the case of non-congested clusters, 12387 cases were correctly classified as non-congested. However, 825 errors were made in misclassifying non-congested situations as congestion, and 4852 errors were made in classifying congested situations as non-congested.

Table 2. Confusion matrix for the city of San Francisco using 45 seconds of forgetting and tolerance of 0.2.

	Congested cells	**Non-congested cells**
Congested clusters	6881	825
Non-congested clusters	4852	12387

The confusion matrix for the city of San Francisco, with a forgetting value of 60 seconds and no tolerance, is presented in Table 3, for congested clusters, 6708 cases were correctly classified as congestion, and for non-congested clusters, 10776 cases were correctly classified as non-congested. However, 1390 errors were made in misclassifying non-congested situations as congestion, and 6293 errors were made in classifying congested situations as non-congested.

Table 3. Confusion matrix for the city of San Francisco using 60 seconds of forgetting and no tolerance.

	Congested cells	**Non-congested cells**
Congested clusters	6708	1390
Non-congested clusters	6293	10776

The confusion matrix for the city of San Francisco, with a forgetting value of 60 seconds and a tolerance of 0.2, is presented in Tabla 4, for congested clusters, 7177 cases were correctly classified as congestion, and for non-congested clusters, 12232 cases were correctly classified as non-congested.However, 921 errors were made in misclassifying non-congested situations as congestion, and 4837 errors were made in classifying congested situations as non-congested.

Table 4. Confusion matrix for the city of San Francisco using 60 seconds of forgetfulness and tolerance of 0.2.

	Congested cells	**Non-congested cells**
Congested clusters	7177	921
Non-congested clusters	4837	12232

These results indicate that, compared to previous parameterizations with a forgetting value of 45 seconds, the model presents slightly better quantities in identifying congested situations. When evaluating the true positive rate, that is, the ability to correctly identify the congested state of traffic, the clusters obtained a high number of matches compared to the congested grid cells in the city of San Francisco.

4. Discussion

In this section, we examine the results obtained from the comparison between the dynamic clustering methodology and the congestion indicator applied to static cells in the city of San Francisco. The evaluation metric is represented by the precision rates which provides an in-depth understanding of the performance of both results.

The precision is obtained using Equation 2:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

 doi:10.20944/preprints202311.1043.v1

where, TP represents the number of true positives, i.e., cases that the model has correctly classified as congested, and FP represents the number of false positives, i.e., cases that the model has incorrectly classified as congested when they are actually non-congested.

In this research, a detailed analysis of the congested cluster classification was carried out in the traffic congestion study, as this classification plays a central role in urban traffic management and in improving mobility in cities. Congested clusters represent problematic traffic scenarios that can have a significant impact on urban mobility and citizens' quality of life. Accurately identifying and classifying these situations is essential for making informed traffic management decisions and applying effective congestion relief strategies.

The precision results for the clusters categorized as congested are displayed in Table 5. The results table provides important insight into the precision of the comparison between the congested clusters of the dynamic cluster method and the congested cells of the static grid method in identifying traffic congestion. Two key parameters, forgetting and tolerance, have been evaluated to understand their impact on the precision of the results.

Table 5. Precision results in congested situations.

Forgetting (seconds)	Tolerance	Precision
45	0	83.48%
45	0.2	89.29%
60	0	82.84%
60	0.2	88.63%

First, forgetting values vary between 45 and 60 seconds, controlling the amount of past information considered in determining congestion. As the forgetting value increases, a longer time window is contemplated, which may influence the precision of the identifications.

The tolerance parameter has been explored at two levels: 0 and 0.2. This tolerance is a value that allows when adjusting the congestion indicator, it tends to slightly increase the precision when allowing for a larger margin of error in the congestion classification.

Looking at the results as a whole, it stands out that the highest precision is achieved when a forgetting value of 45 seconds is used in combination with a tolerance of 0.2, resulting in an precision of 89.29%. In addition, a forgetting of 60 seconds together with a tolerance of 0.2 also provides a high level of precision, reaching a value of 88.63%.

The analysis reveals that as forgetting decreases and tolerance increases, precision tends to improve. This suggests that reducing the amount of past information considered and allowing a larger margin of error in congestion classification would result in higher traffic congestion prediction precision.

The results highlight the crucial influence of the parameters forgetfulness and tolerance on the precision of traffic congestion prediction. It is clear that the appropriate choice of these values is a determining factor in achieving optimal precision in the classification of congestion situations. The configuration of these parameters must be precisely aligned with the specific application requirements and prediction objectives.

However, it is important to note that this improvement in precision by reducing the forgetting value and increasing the tolerance can also have implications for other aspects of the analysis. A lower forgetting value means that a narrower time window is being considered, which may result in the loss of relevant information in the long term. In addition, a higher tolerance implies a wider margin of error, which could allow the inclusion of noisy data that affects precision in certain situations.

Therefore, finding the right balance between these parameters is a key challenge in the practical application of these methods. The choice of optimal values for forgetting and tolerance will depend on the specific needs of the congestion prediction task and the importance of maintaining precision

 doi:10.20944/preprints202311.1043.v1

compared to other factors, such as retention of historical information and management of noise in the data.

In analyzing these results, it is essential to highlight the efficiency of the clustering algorithm in detecting vehicle congestion compared to the method based on static cells in fixed regions. The high levels of precision strengthen the algorithm's ability to identify congestion patterns in the data and anticipate future situations.

Furthermore, looking at the comparison made in the model testing, it is evident that the lack of adaptability of the congestion indicator applied to static cells to adjust to the evolution of the data and changes in the distribution of clusters may affect the quality of the results. If congestion data and traffic flows are not properly identified due to this lack of adaptability, the detection of congested areas may lack reliability.

On the other hand, dynamic clustering methodology that accounts for variations in data flows and adapts to changes in cluster distribution provides an accurate representation of traffic dynamics. As shown in Table 6, statistical data for a specific cluster indicate that the speeds of cycle 4, although belonging to the same cluster, have undergone unusual changes. This is due to the fact that the cluster incorporates information from different registered vehicles, which allows better adaptation to traffic evolution.

By continuously recalibrating the centroid position and adjusting the hyperbox based on evolving data, this methodology effectively captures variations in densities and shapes of the road segments, this can be seen in Table 7 which shows the information used to analyze the Cycle 5 road segments, in this table it can be seen that out of four segments identified only three of them have recorded vehicles, as each segment is analyzed independently it is possible for a vehicle traveling through several segments to be counted as a single vehicle in the context of another segment, the visual representation for this table is associated with Figure 10c. This allows groupings to be made based on up-to-date data and realistic evaluations.

Table 6. Example of the evolution of a cluster.

Cicle	Points	Vehicles	Speed
1	4	2	17.94
2	3	2	17.91
3	2	1	16.43
4	3	1	3.47
5	9	3	16.96
6	6	3	14.32

Table 7. Example of the dynamism of road segments in a cluster in cycle 5.

Segment	Vehicles	Length (meters)	Density	Congestion indicator
1	3	13.73	0.218	2.14
2	2	14.29	0.139	1.39
3	1	7.87	0.126	1.89
4	0	2.14	0.000	N/A
Total	**6**	**35.89**		

5. Conclusions

The obtained results highlight the effectiveness of the dynamic clustering methodology compared to the static cell-based method for classifying congestion conditions. By allowing dynamic clustering of vehicle trajectory data and performing specific analysis for each cluster, this methodology facilitates early and accurate detection of congested traffic problem areas.

The application of dynamic clustering methods presents itself as a highly promising strategy. These methods have the ability to adapt to constant changes in urban traffic, capturing constantly evolving mobility patterns. The relevance of the forgetting factor lies in its ability to keep the clusters up-to-date, considering both recent and old locations. This ensures that the clusters accurately reflect current traffic dynamics, allowing emerging congestion to be identified early.

Carefully tuning the forgetting and tolerance parameters is critical to obtain accurate results in the comparison between dynamic cluster and static grid methods in traffic congestion prediction. These findings are essential for improving the effectiveness of classification models in traffic management applications.

The methodology based on dynamic clustering stands out for its adaptability to changes in traffic, providing a complete and up-to-date view of vehicular behavior in urban areas. These results support the effectiveness of clusters as a valuable tool for improving traffic management and reducing congestion problems in cities.

As for future research, it is essential to explore in depth the possible reasons behind the observed decrease in precision in the congested classification. In addition, it is proposed to conduct experiments in larger areas, under extreme congestion conditions and in highly complex traffic scenarios. These experiments will provide additional insight into the limits and robustness of the proposed methodology.

Author Contributions: Conceptualization, G.R.; methodology, G.R.; software, G.R.; validation, L.L and C.E.; formal analysis, L.L and C.E.; data curation, G.R. supervision, R.T., A.B. and J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The GPS trajectory dataset analyzed in this study can be found here: https://github.com/gary-reyes-zambrano/SanFrancisco-trajectory-dataset.

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

References

1. Soumia Goumiri, S. & Djahel, S. Smart Mobility in Smart Cities: Emerging Challenges, Recent Advances and Future Directions. *Journal Of Intelligent Transportation Systems* **2023**, *0*, 1-37, doi:10.3390/ijgi9120721.
2. Sepehr G. Dehkordi & Glaser, S. Including Network Level Safety Measures in Eco-Routing. *Journal Of Intelligent Transportation Systems* **2022**, *0*, 1-14, doi:10.1080/15472450.2022.2129022.
3. Wen, Z. & Weng, X. Inferring the Number of Vehicles between Trajectory-Observed Vehicles. *Journal Of Intelligent Transportation Systems* **2023**, *0*, 1-14, doi:10.1080/15472450.2023.2227940.
4. Matej Cebecauer, D. & Burghout, W. Revealing Representative Day-Types in Transport Networks Using Traffic Data Clustering. *Journal Of Intelligent Transportation Systems* **2023**, *0*, 1-24, doi:10.1080/15472450.2023.2205020.
5. Vishal C. Kummetha & Dokur, O. Proactive Congestion Management via Data-Driven Methods and Connected Vehicle-Based Microsimulation. *Journal Of Intelligent Transportation Systems* **2022**, *0*, 1-17, doi:10.1080/15472450.2022.2140047.
6. Maiti, N. & Chilukuri, B. Estimation of Local Traffic Conditions Using Wi-Fi Sensor Technology. *Journal Of Intelligent Transportation Systems* **2023**, *0*, 1-18, doi:10.1080/15472450.2023.2177103.
7. Heshami, S. & Kattan, L. A Stochastic Microscopic Based Freeway Traffic State and Spatial-Temporal Pattern Prediction in a Connected Vehicle Environment. *Journal Of Intelligent Transportation Systems* **2023**, *0*, 1-27, doi:10.1080/15472450.2022.2130291.
8. Li, L., Jiang, R., He, Z., Chen, X. & Zhou, X. Trajectory Data-Based Traffic Flow Studies: A Revisit. *Transportation Research Part C: Emerging Technologies*, **2020**, *114*, 225-240, doi:10.1016/j.trc.2020.02.016.
9. Jain, A. Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters* **2009**.
10. Tork, H. Spatio-temporal clustering methods classification. In *Doctoral Symposium On Informatics Engineering*, Faculdade de Engenharia da Universidade do Porto Porto, Portugal, **2012**, *1*, 199-209.

11. Mazimpaka, J. & Timpf, S. Trajectory data mining: A review of methods and applications. *Journal Of Spatial Information Science* **2016**, *2016*, 61-99.
12. Han, J., Kamber, M. & Tung, A. Spatial clustering methods in data mining. *Geographic Data Mining And Knowledge Discovery* **2001**, *0*, 188-217.
13. Zeng, J., Xiong, Y., Liu, F., Ye, J. & Tang, J. Uncovering the Spatiotemporal Patterns of Traffic Congestion from Large-Scale Trajectory Data: A Complex Network Approach. *Physica A: Statistical Mechanics And Its Applications* **2022**, *604*, 127871, doi:10.1016/j.physa.2022.127871.
14. Ester, M., Kriegel, H., Sander, J., Xu, X. & Others A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* **1996**, *96*, 226-231.
15. Zhang, H. & Yang, J. A Case Retrieval Strategy for Traffic Congestion Based on Cluster Analysis. *Mathematical Problems In Engineering* **2022**, *2022*, 1-8, doi:10.1155/2022/5234230.
16. Lee, J., Han, J. & Whang, K. Trajectory clustering: a partition-and-group framework. In *Proceedings Of The 2007 ACM SIGMOD International Conference On Management Of Data - SIGMOD '07* **2007**, *0*, 593, doi:10.1145/1247480.1247546.
17. Mao, Y., Zhong, H., Qi, H., Ping, P. & Li, X. An Adaptive Trajectory Clustering Method Based on Grid and Density in Mobile Pattern Analysis. *Sensors* **2017**, *17*, 2013, doi:10.3390/s17092013.
18. Liu, Y., Yan, X., Wang, Y., Yang, Z. & Wu, J. Grid Mapping for Spatial Pattern Analyses of Recurrent Urban Traffic Congestion Based on Taxi GPS Sensing Data. *Sustainability* **2017**, *9*, 533, doi:10.3390/su9040533.
19. Lou, J. & Cheng, A. Detecting Pattern Changes in Individual Travel Behavior from Vehicle GPS/GNSS Data. *Sensors* **2020**, *20*, doi:10.3390/s20082295.
20. Saeedmanesh, M. & Geroliminis, N. Dynamic Clustering and Propagation of Congestion in Heterogeneously Congested Urban Traffic Networks. *Transportation Research Part B: Methodological* **2017**, *105*, 193-211, doi:10.1016/j.trb.2017.08.021.
21. Makara, L., Maric, P. & Pekar, A. Public Transport Congestion Detection Using Incremental Learning. *Pervasive And Mobile Computing* **2023**, *91*, 101769, doi:10.1016/j.pmcj.2023.101769.
22. Sun, S., Chen, J. & Sun, J. Traffic congestion prediction based on GPS trajectory data. *International Journal Of Distributed Sensor Networks* **2019**, *15*, doi:10.1177/1550147719847440.
23. Bratsas, C., Koupidis, K., Salanova, J., Giannakopoulos, K., Kaloudis, A. & Aifadopoulou, G. A Comparison of Machine Learning Methods for the Prediction of Traffic Speed in Urban Places. *Sustainability* **2020**, *12*, 142, doi:10.3390/su12010142.
24. Kamble, S. & Kounte, M. Machine Learning Approach on Traffic Congestion Monitoring System in Internet of Vehicles. *Procedia Computer Science* **2020**, *171*, 2235-2241, doi:10.1016/j.procs.2020.04.241.
25. Cherkaoui, B., Beni-Hssane, A., Fissaoui, M. & Erritali, M. Road Traffic Congestion Detection in VANET Networks. *Procedia Computer Science* **2019**, *151*, 1158-1163, doi:10.1016/j.procs.2019.04.165.
26. Luo, P., Liu, Y., Wang, Z., Chu, J. & Yang, G. A Novel Congestion Control Algorithm Based on Inverse Reinforcement Learning with Parallel Training. *Computer Networks* **2023**, *237*, 110071, doi:10.1016/j.comnet.2023.110071.
27. Zulfikar, M. & Suharjito Detection Traffic Congestion Based on Twitter Data Using Machine Learning. *Procedia Computer Science* **2019**, *157*, 118-124, doi:10.1016/j.procs.2019.08.148.
28. Erdelić, T., Carić, T., Erdelić, M., Tišljarić, L., Turković, A. & Jelušić, N. Estimating congestion zones and travel time indexes based on the floating car data. *Computers, Environment And Urban Systems* **2021**, *87*, 101604, doi:10.1016/j.compenvurbsys.2021.101604.
29. Boarnet, M., Kim, E. & Parkany, E. Measuring Traffic Congestion. *Transportation Research Record: Journal Of The Transportation Research Board* **1998**, *1634*, 93-99, doi:10.3141/1634-12.
30. Pei, Y., Cai, X., Li, J., Song, K. & Liu, R. Method for Identifying the Traffic Congestion Situation of the Main Road in Cold-Climate Cities Based on the Clustering Analysis Algorithm. *Sustainability* **2021**, *13*, 9741, doi:10.3390/su13179741.
31. Seong, J., Kim, Y., Goh, H., Kim, H. & Stanescu, A. Measuring Traffic Congestion with Novel Metrics: A Case Study of Six U.S. Metropolitan Areas. *ISPRS International Journal Of Geo-Information* **2023**, *12*, 130, doi:10.3390/ijgi12030130.
32. Zang, J., Jiao, P., Liu, S., Zhang, X., Song, G. & Yu, L. Identifying Traffic Congestion Patterns of Urban Road Network Based on Traffic Performance Index. *Sustainability* **2023**, *15*, 948, doi:10.3390/su15020948.

33. Zhao, X., Hu, L., Wang, X. & Wu, J. Study on Identification and Prevention of Traffic Congestion Zones Considering Resilience-Vulnerability of Urban Transportation Systems. *Sustainability* **2022**, *14*, 16907, doi:10.3390/su142416907.
34. Azimi, M. & Zhang, Y. Categorizing Freeway Flow Conditions by Using Clustering Methods. *Transportation Research Record: Journal Of The Transportation Research Board* **2010**, *2173*, 105-114, doi:10.3141/2173-13.
35. Rempe, F., Huber, G. & Bogenberger, K. Spatio-Temporal Congestion Patterns in Urban Traffic Networks. *Transportation Research Procedia* **2016**, *15*, 513-524, doi:10.1016/j.trpro.2016.06.043.
36. Shang, Q., Yu, Y. & Xie, T. A Hybrid Method for Traffic State Classification Using K-Medoids Clustering and Self-Tuning Spectral Clustering. *Sustainability* **2022**, *14*, 11068, doi:10.3390/su141711068.
37. Zhang, Y., Ye, N., Wang, R. & Malekian, R. A Method for Traffic Congestion Clustering Judgment Based on Grey Relational Analysis. *ISPRS International Journal Of Geo-Information* **2016**, *5*, doi:10.3390/ijgi5050071.
38. Kim, J. & Mahmassani, H. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. *Transportation Research Procedia* **2015**, *9*, 164-184.
39. Almeida, A., Brás, S., Sargento, S. & Oliveira, I. Exploring Bus Tracking Data to Characterize Urban Traffic Congestion. *Journal Of Urban Mobility* **2023**, *4*, 100065, doi:10.1016/j.urbmob.2023.100065.
40. Reyes, G., Lanzarini, L., Estrebou, C. & Fernandez Bariviera, A. Dynamic grouping of vehicle trajectories. *Journal Of Computer Science And Technology* **2022**, *22*, e11, doi:10.24215/16666038.22.e11.
41. Gao, H., Yan, Z., Hu, X., Yu, Z., Luo, W., Yuan, L. & Zhang, J. A Method for Exploring and Analyzing Spatiotemporal Patterns of Traffic Congestion in Expressway Networks Based on Origin–Destination Data. *ISPRS International Journal Of Geo-Information* **2021**, *10*, doi:10.3390/ijgi10050288.
42. Nguyen, D., Dow, C. & Hwang, S. An Efficient Traffic Congestion Monitoring System on Internet of Vehicles. *Wireless Communications And Mobile Computing* **2018**, *2018*, 1-17, doi:10.1155/2018/9136813.