ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Reyes, Gary; Lanzarini, Laura; Estrebou, Cesar; Bariviera, Aurelio F.

Article — Published Version

Dynamic grouping of vehicle trajectories

Journal of Computer Science & Technology

Suggested Citation: Reyes, Gary; Lanzarini, Laura; Estrebou, Cesar; Bariviera, Aurelio F. (2022) : Dynamic grouping of vehicle trajectories, Journal of Computer Science & Technology, ISSN 1666-6038, Universidad Nacional de La Plata, La Plata, Vol. 22, Iss. 2, pp. 141-150, https://doi.org/10.24215/16666038.22.e11 , https://journal.info.unlp.edu.ar/JCST/article/view/2059

This Version is available at:
https://hdl.handle.net/10419/268416

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

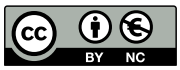
Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



http://creativecommons.org/licenses/by-nc/4.0

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

- ORIGINAL ARTICLE -

Dynamic grouping of vehicle trajectories

Agrupamiento dinámico de trayectorias vehiculares

Gary Reyes[1], Laura Lanzarini[2], César Estrebou[2], and Aurelio Bariviera[3]

[1]*Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil,Cdla. Universitaria Salvador Allende, Guayaquil 090514, Ecuador*
{gary.reyesz}@ug.edu.ec
[2]*Universidad Nacional de La Plata, Facultad de Informática,Instituto de Investigación en Informática LIDI (Centro CICPBA) 1900 La Plata, Buenos Aires, Argentina*
{laural, cesarest}@lidi.info.unlp.edu.ar
[3]*Universitat Rovira i Virgili, Department of Business, Reus, Spain*
{aurelio.fernandez}@urv.cat

Abstract

Vehicular traffic volume in large cities has increased in recent years, causing mobility problems; therefore, the analysis of vehicle flow data becomes a relevant research topic. Intelligent Transportation Systems monitor and control vehicular movements by collecting GPS trajectories, which provides the geographic location of vehicles in real time. Thus information is processed using clustering techniques to identify vehicular flow patterns. This work presents a methodology capable of analyzing the vehicular flow in a given area, identifying speed ranges and keeping an interactive map updated that facilitates the identification of possible traffic jam areas. The obtained results on three data sets from the cities of Guayaquil-Ecuador, Rome-Italy and Beijing-China are satisfactory and clearly represent the speed of movement of the vehicles, automatically identifying the most representative ranges in real time.

Keywords: Dynamic clustering, data stream, vehicular trajectories

Resumen

El volumen de tráfico vehicular de las grandes ciudades se ha incrementado en los últimos años originando problemas de movilidad, por ello el análisis de los datos del flujo vehicular toma importancia para los investigadores. Los Sistemas Inteligentes de transportación realizan el monitoreo y control vehicular recolectando trayectorias GPS, información que brinda en tiempo real la ubicación geográfica de los vehículos. Su procesamiento por medio de técnicas de agrupamiento permite identificar patrones sobre el flujo vehicular. Este trabajo presenta una metodología capaz de analizar el flujo vehicular en un área dada, identificando los rangos de velocidades y manteniendo actualizado un mapa interactivo que facilita la identificación de zonas de posibles atascos. Los resultados obtenidos sobre tres conjuntos de datos de las ciudades de Guayaquil-Ecuador, Roma-Italia y Beijing-China son satisfactorios y representan claramente la velocidad de desplazamiento de los vehículos identificando de manera automática los rangos más representativos para cada instante de tiempo.

Palabras claves: Agrupamiento dinámico, flujo de datos, trayectorias vehiculares

1 Introduction

Nowadays, the constant increase in traffic volume in large cities causes problems in vehicular flow and, therefore, the analysis of the generated data by vehicle monitoring and control systems becomes relevant. Using descriptive techniques, relationships between vehicular trajectories can be identified, which facilitates vehicular flow analysis. These currently provide solutions in very diverse areas, such as health, finance, telecommunications, agriculture and transportation, among others [1].

Data clustering is a widely used technique to identify common features between instances of the same problem [2]. Over time, researchers have proposed improvements to the identified limitations in some techniques, such as those carried out by Bahmani et al. [3], where a correct initialization of the algorithm is achieved in a much shorter time. In other cases, the techniques have been adapted to work in a specific context, for example, for spatial data mining [4, 5, 6] or for GPS trajectory analysis [7]. A GPS trajectory is defined by a set of geographic locations, each of which represented by its latitude and longitude, at an instant of time. This work proposes a methodology for analyzing vehicular flow in traffic through the analysis of GPS trajectories. To do this, each zone within an area of interest is characterized based on the average speed and the number of vehicles it contains in a given period of time. Zones are delimited at the beginning of the process, and their size depends on the

accuracy with which the analysis is to be carried out. Then, using a dynamic clustering technique, areas with similar characteristics are identified and an interactive map is built on which the speed ranges corresponding to the current vehicular flow and the areas where they occur can be observed. This methodology can be used, along with other tools, by city traffic managers to plan urban roads, detect critical points in traffic flow, identify anomalous situations, predict future mobility behaviors, analyze vehicular flow, and so forth. The proposed methodology was used to characterize the data corresponding to generated GPS trajectories by a group of students from the University of Guayaquil in Ecuador, historical data of taxis from the city of Rome in Italy, and GPS trajectories from the city of Beijing in China. The obtained results allowed identifying, in each city, different instants of time where vehicles have the same speeds.

This article is organized as follows: Section 2 analyzes some related works that were identified within the literature and present various solutions to the problem, Section 3 describes the proposed methodology, Section 4 discusses the obtained results, and Section 5 presents the conclusions and future lines of work.

2 Related Work

Clustering techniques have been used in trajectory analysis for several years. In general, these are adaptations of conventional algorithms using similarity metrics specially designed for trajectories [8, 9].

Trajectory clustering has certain characteristics that should be specially considered. Firstly, the way in which the data that makes up the trajectory will be entered should be defined. Doing it individually, processing the locations one by one, makes identifying similar routes harder, which is why it is usually segmented beforehand. Such is the case of these authors [10, 11, 12], where the authors propose using segmentation in a stage prior to clustering. This allows improving algorithm performance and yields better quality clusters. Secondly, the used metric to compare trajectory segments determines the result of the clustering. Various examples can be found in the literature. For example, the Tra-DBScan algorithm [13], which adapts the well-known DBScan algorithm [14] by adding a trajectory segmentation phase, partitions trajectories into sections and uses the Hausdorff distance as a measure of similarity. Also Reyes et al. [15] present a GPS vehicle trajectory clustering method that uses angular information to segment the trajectories and a similarity metric guided by a pivot. Yu et al. [16] propose an improved trajectory model and a new clustering algorithm is presented, with a similarity measure that calculates the distance between two trajectories based on multiple data characteristics, maximizing the similarity between them.

Another important characteristic that should be taken into account when applying dynamic clustering techniques is how cluster centers are represented. This is the case of the Improved DBScan algorithm [17], which improves the traditional DBScan algorithm using a proprietary density measurement method that suggests using the new concept of movement capacity and introduces the data field theory. Similarly, Ferreira et al. [18] presented a new trajectory clustering technique that uses vector fields to represent cluster centers and propose a definition of trajectory similarity. Research efforts in this area continue to date, as demonstrated by other authors [19, 20].

In summary, it can be stated that clustering techniques have shown to perform well in the analysis of vehicle trajectories, although their parameterization remains an interesting challenge. This is due to the fact that these are unsupervised techniques that generally combine distance and density metrics to control cluster creation.

In this article, a dynamic clustering algorithm has been used for data stream. This type of algorithm processes data streams to overcome some of the limitations of traditional clustering algorithms, which tend to iterate over the dataset more than once, causing greater memory use and increasing execution time [21, 22]. As the data distribution of each stream changes continuously, it is important that these clustering algorithms that process data streams generate dynamic clusters, where the number of clusters will depend on stream data distribution [23, 24].

The used methodology in this work was previously defined defined by the authors of this work [25]. This methodology proposes segmenting trajectories and generating a new representation based on the information to be grouped. As an interpretation tool for the user, the use of interactive automatic maps is incorporated. This helps visualize the result of the clustering process. In article proposed by Reyes et al. [25], the proposed methodology was used to analyze defined trajectories on the city of Guayaquil-Ecuador with the aim of automatically obtaining frequent speed ranges for different time intervals.

In this work, more tests were carried out, and the results of applying this methodology to analyze trajectories corresponding to two new cities were measured: Rome (Italy) and Beijing (China).

In addition, dynamic maps were improved by adding a new range of colors to facilitate the identification of low and high speeds. Information regarding the density of the groups was also used to determine, in a more stable way, the ranges of higher or extremely slow speeds.

3 Methodology

This work proposes a methodology for analyzing vehicular flow in traffic. This methodology is represented in Figure 1, and it has three steps. The first step con-

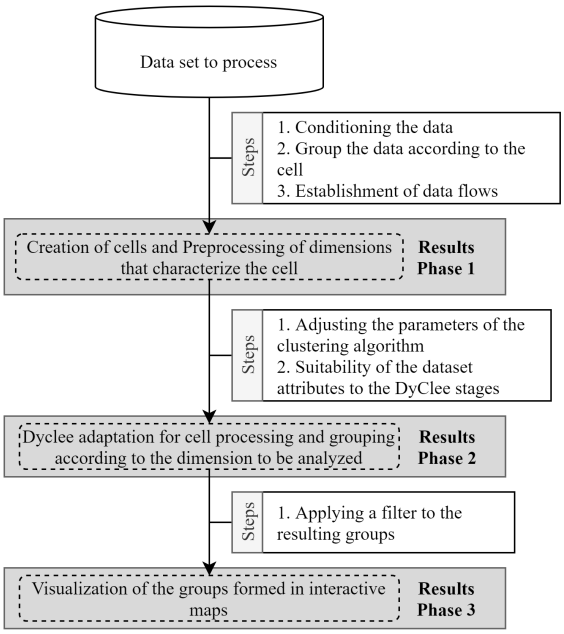


Figure 1: Proposed methodology

sists of adequately representing trajectory data within the area of interest; the second step uses a proposed dynamic clustering algorithm to identify relationships, and the third step consists of creating interactive resources for visualizing the results. Each of these steps is described below.

3.1 Vehicular Flow Representation

The structure of the records to be processed contains information on longitude, latitude, time, speed and the id of the trajectory. It has been established that the speed ranges used for the grouping are within 20 km/h.

The first step is providing an adequate representation of the data that makes up the trajectories. To do this, the area of interest must be defined first, indicating the geographical area to which the trajectories to be analyzed belong. Once the area is set, it is split into uniform cells, or smaller zones. The size of each cell will depend on the accuracy with which vehicular flow is to be analyzed. In this work, cells of 200x200 meters were used.

This is a significant aspect that should not be disregarded, since the information to be analyzed corresponds to a summary of what is happening in each cell during a given period of time. The proposed methodology here consists of analyzing what is happening in each cell as a whole instead of considering each vehicle trajectory separately. This simplifies both analysis and visualization.

In particular, in this article, the data corresponding to the vehicular flow, represented in each cell, were analyzed in batch mode in periods of equal duration. The duration of each period is a parameter of the algorithm that must be set a priori.

Table 1: Associated concepts with DyClee processing used in the proposed method

Element	Definition
Microcluster	This represents the dataset with similar characteristics.
HiperBox	The area of the microcluster.
Relative Size	Size of the microclusters compared to the processed area..

Each period is considered an evolution, since the clustering result is updated with the incorporation of each block of data sequentially.

In each evolution, a data stream enters and the corresponding calculations are carried out for each cell, obtaining characteristic information.

3.2 Proposed Algorithm

In the second step, a proposed algorithm based on the methodology defined by Reyes et al. [25] is used to process the trajectories within each cell.

Table 1 identifies the basic concepts that were used in the proposed method.

As previously stated, the first of the two stages of the DyClee algorithm is used for building microclusters. In this article, instead of directly using GPS locations and their density, as proposed in the original article, the trajectory speeds of the sections included in each cell were considered. Thus, for a given period of time, each cell will be represented by the average speed of the sections of trajectories it contains. This involves splitting the trajectories appropriately considering that speed variations can lead to a vehicle traveling at very high speed not being registered (or having very few GPS locations) when passing through a cell. In addition, speeds should be averaged considering the vehicles and not the number of registered locations. Regarding microcluster size, the value of the ”Relative Size” parameter specifies the relative size of the ”HyperBox” parameter with respect to the area to be processed. In other words, as the value decreases, the number of microclusters increases, and vice versa. The cells of each evolution are clustered in a distance-based stage that considers the smallest existing speed difference between the cluster and the analyzed cell for the cell to be part of that cluster and in which case the information of this cluster is updated; if no nearby group exists, a new microcluster will be formed with the analyzed cell.

In its second stage, the algorithm analyzes microcluster densities and classifies them into two categories — dense and semi-dense. As a first step, those dense microclusters that are connected directly are joined. Two microclusters are considered to be connected directly if the maximum distance at which the centroids of two microclusters with similar densities can be

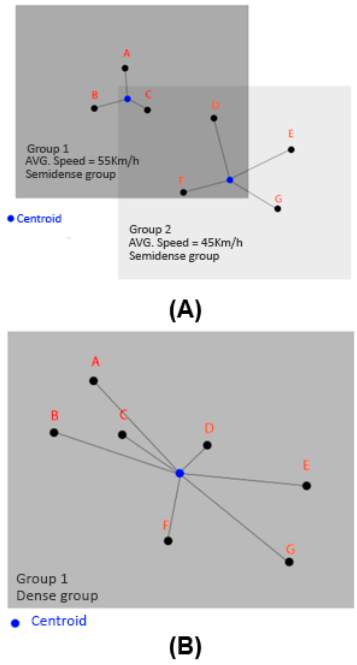


Figure 2: Second stage of proposed algorithm. (A) Identification of microclusters that are connected directly, and (B) Resulting expansion

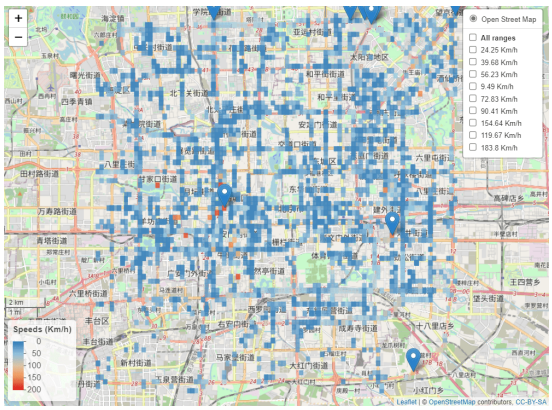


Figure 3: Interactive map showing the evolution with the information of the selected clusters

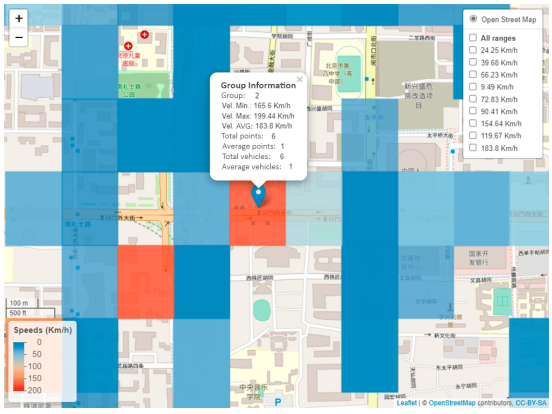


Figure 4: Available information using markers in an interactive map representing an evolution.

found does not exceed the value of the "HyperBox" parameter. In other words, as the value of the "HyperBox" parameter increases, the number of clusters will be less and, therefore, the speed ranges in the area of interest will have greater amplitude. Figure 2 illustrates the operation of this stage. During these processes, information regarding the performance of the stages is recorded separately.

Then the results are then exported and the information of the cells that make up each group is saved. Additionally, a summary of the results of the groups in the execution performed is stored.

Before going on to the visualization of the results, the low density clusters are processed, where all the summaries of the clusters are used, they are sorted according to their speed, then the clusters with a single GPS point are excluded and which in turn are found without more contiguous low density clusters. Those low density clusters that are found with more contiguous low density clusters are selected to be considered as a single group, this union is calculated weighted according to the number of cells to determine the average speed of the new group and the information of number of points, cells and vehicles is totaled.

3.3 Clustering Visualization

As previously mentioned, trajectory information is analyzed at 3-minutes intervals.

With the result of each clustering, an interactive map is created in which the relevant information of each cluster for a specific development can be analyzed graphically and dynamically. In this work, each cluster is represented on the map with its respective speed ranges, accompanied by a color scale chosen in a rigorous manner for easy distinction of the colors, in this scale, the shades of red colors represent the highest speeds and the shades of blue colors represent the lowest speeds. Since speed limits may vary from city to city, we have chosen to use a speed scale relative to the city under analysis.

On the other hand, since the result of the clustering process carried out is recorded, all maps can be rebuilt from the start of the vehicular flow analysis. This provides a quick view of traffic status. Each map's interface includes layer selection controls and reference legends to be able to read the represented results. Therefore, two types of maps van be generated – a map of the last evolution or period of time analyzed, and a map with all the evolutions.

In the case of the map corresponding to a particular evolution, each cluster has been represented on a different layer, and the user can select one or several layers to filter the information. As shown in Figure 4, a color scale has been used based on the average speed for each group; this scale goes from green, representing lower speeds, to red, representing higher speeds. The map also allows displaying markers that show information on both the group and the selected cell. Figure

Figure 5: Guayaquil area and dataset

4 illustrates the latter.

4 Results and Discussion

4.1 Used data

To verify that operation of the proposed methodology in this article, a set of own data collected in the city of Guayaquil and two databases of public trajectories extracted from repositories corresponding to the cities of Rome and Beijing were used. A brief description of each dataset is given below:

Guayaquil dataset

This dataset was collected on October 28, 2017 and includes 218 trajectories made by university students who travel in some form of transport such as taxi, motorcycle, subway, etc. The locations in this dataset were collected by smartphones with an average time interval of 5 seconds between two consecutive locations. Each record contains the following data: trajectory id, latitude, longitude, time, username, email and type of transport.

Since this is a reduced set of trajectories, the analysis was carried out for times between 4:30 p.m. and 6:30 p.m., as most records were within this period. As a result of this filtering process, 30,557 records were obtained, representing 206 trajectories of the entire dataset. The area that represents the selected dataset is shown in Figure 5.

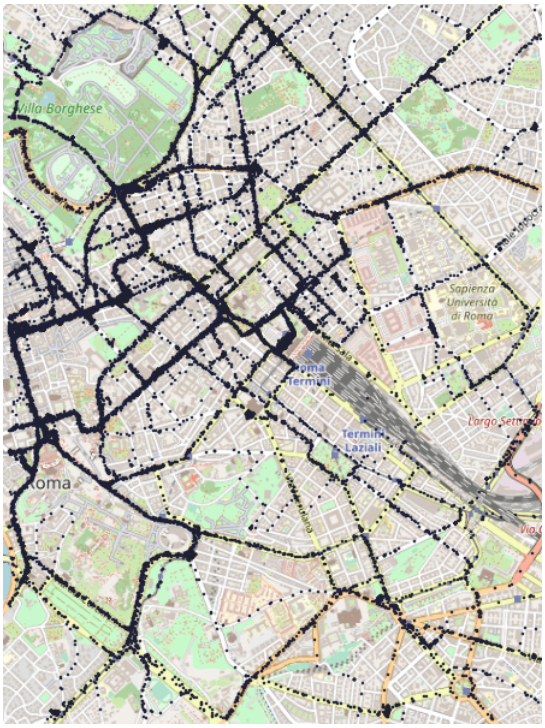


Figure 6: Rome area and dataset

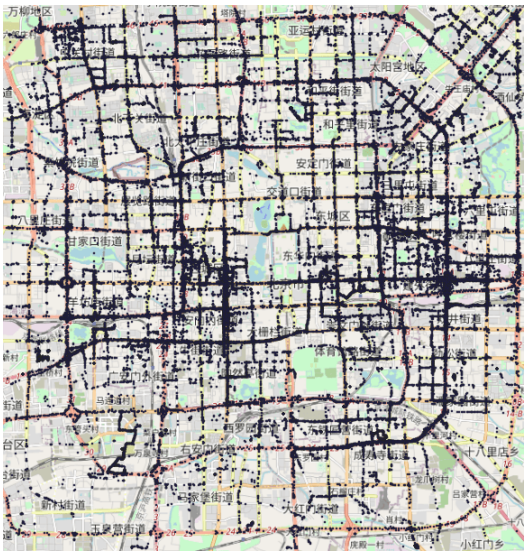


Figure 7: Beijing area and dataset

Rome dataset

The dataset for the city of Rome[1] was collected on February 12, 2014; it contains 137 trajectories recorded by taxis using GPS devices, with an average time interval of 10 seconds between two consecutive locations. Each record contains the following data: trajectory id, latitude, longitude, time, speed, and direction.

For this second set of trajectories, the analysis included all trajectories recorded between 6:00 p.m. and 8:00 p.m. As a result of this filtering process, 33,793 records were obtained, representing 137 trajectories of the entire dataset. Figure 6 illustrates the analyzed area and the locations of these trajectories.

Beijing dataset

The third public dataset was recorded on February 3, 2008 in the city of Beijing [2]. These data correspond to 630 taxi trajectories. Vehicle locations were collected by GPS devices with an average time interval of 5 seconds between two consecutive locations. Each record contains the following data. trajectory id, latitude, longitude, and time. Speed information was calculated for this article.

For this third set of trajectories, the analysis included all recorded trajectories between 02:30 a.m. and 04:00 a.m. As a result of this filtering process, 58,745 records were obtained, representing 630 trajectories of the entire dataset. The area that represents the selected dataset is shown in Figure 7.

4.2 Obtained results

For each dataset, the relevant time period was analyzed in blocks of 15 minutes, which resulted in a total of 8 blocks in the case of Guayaquil and Rome, and only 6 for Beijing. In each case, blocks were analyzed consecutively. This 15-minute time period could be considered excessive, since its duration is in relation to the volume of data available.

Then, to use the DyClee algorithm, values had to be defined for the "Relative size" and "HyperBox" parameters. "Relative Size" was set at 0.2 for the Guayaquil dataset, using limits between 0 and 100 km/h, and it was set at 0.1 for the Rome and Beijing datasets, using limits between 0 and 200 km/h. These values were established so as to have homogeneous ranges between the experiments, considering that the Rome and Beijing datasets include speeds above 100 km/h. Based on this value and on the processing area, the value of the "HyperBox" parameter was set at 20 km/h.

Figures 8 , 9 and 10 show the obtained maps as a result of the evolutions carried out for each dataset. Below each map, the information corresponding to the obtained cluster with the adaptation of the DyClee algorithm is indicated. For each group, its minimum, maximum and average speeds are indicated, as well as any speed deviations. In each map, green represents the lower speed range, and red represents the higher speed range.

As it can be seen form the maps depicted in Figure 8 (Guayaquil), large changes can be noted in the dispersion of the cells in the different groups, with the initial and final evolutions being the ones that present the most concentrated groups. This is due to the fact that the data correspond to recorded trajectories by students of the University of Guayaquil at specific times. On the other hand, cell distribution is uniform throughout the evolutions in the maps corresponding to the data for Rome and Beijing (figures 9 and 10) because they correspond to regular trajectories.

In Figure 10, depending on the range of used colors, it can be seen that, for all six evolutions, relatively high speed ranges are found in the outer northeast zone, despite the fact that the analyzed data were recorded between 02:30 a.m. and 04:00 a.m., which is a time with low vehicular traffic. Additionally, between 03:00 a.m. and 03:15 a.m., an area with high speeds is identified in the southeast sector of the city. Lower-end velocities are constant during the 6 evolutions, with a concentration in the east-central zone.

In addition, for the 3 datasets, speed deviations in the clusters formed in each evolution remain within a range between 3 and 5 km/h, which is low for the speed intervals covered by each cluster.

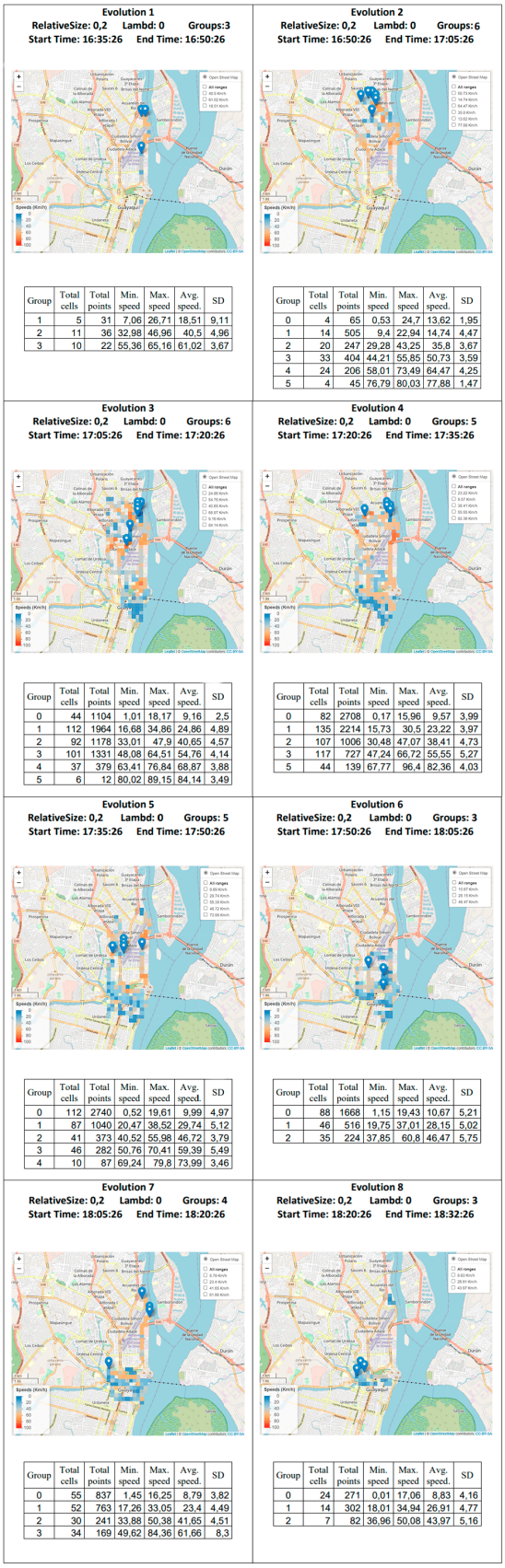
In Figures 11, 12 and 13, series graphs were used to analyze average speed variation in greater detail for each cluster in each evolution. In all of them, the horizontal axis corresponds to the time label to which the trajectory segments analyzed in each evolution belong, and the vertical axis represents the average speed of each cluster.

Figure 11, which represents Guayaquil average speeds, shows that the three lowest speed ranges are present in all the evolutions, while the highest speeds are concentrated mainly in the first half of the analyzed period. This is indicative of the speed variation that occurs in vehicular flow over time, where traffic flows at high speeds during the first hour of the analysis, and lower speeds are observed in the last 45 minutes.

As regards the Rome dataset, Figure 12 shows that there are three speed ranges in all evolutions. Both in this set of trajectories and in those from Guayaquil, the maximum speeds of these stable ranges are below 50 km/h. Similarly, Figure 13 shows that the speed ranges corresponding to the taxi routes in Beijing throughout the entire analysis reach 70 km/h.

[1]Rome dataset available at `https://github.com/gary-reyes-zambrano/Roma-Dataset`

[2]Beijing dataset available at `https://github.com/gary-reyes-zambrano/Beijing-Dataset`



Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	5	31	7,06	26,71	18,51	9,11
2	11	36	32,98	46,96	40,5	4,96
3	10	22	55,36	65,16	61,02	3,67

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
0	4	65	0,53	24,7	13,62	1,95
1	14	505	9,4	22,94	14,74	4,47
2	20	247	29,28	43,25	35,8	3,67
3	33	404	44,21	55,85	50,73	3,59
4	24	206	58,01	73,49	64,47	4,25
5	4	45	76,79	80,03	77,88	1,47

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
0	44	1104	1,01	18,17	9,16	2,5
1	112	1964	16,68	34,86	24,86	4,89
2	92	1178	33,01	47,9	40,65	4,57
3	101	1331	48,08	64,51	54,76	4,14
4	37	379	63,41	76,84	68,87	3,88
5	6	12	80,02	89,15	84,14	3,49

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
0	82	2708	0,17	15,96	9,57	3,99
1	135	2214	15,73	30,5	23,22	3,97
2	107	1006	30,48	47,07	38,41	4,73
3	117	727	47,24	66,72	55,55	5,27
5	44	139	67,77	96,4	82,36	4,03

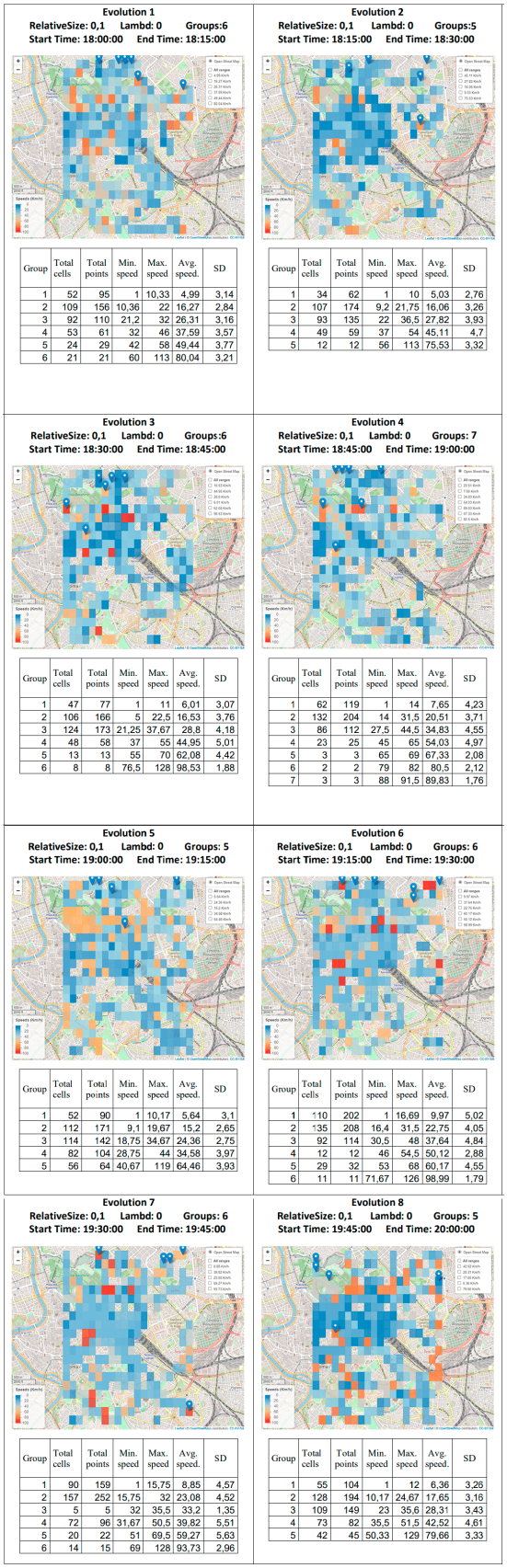
Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
0	112	2740	0,52	19,61	9,99	4,97
1	87	1040	20,47	38,52	29,74	5,12
2	41	373	40,52	55,98	46,72	3,79
3	46	282	50,76	70,41	59,39	5,49
4	10	87	69,24	79,8	73,99	3,46

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
0	88	1668	1,15	19,43	10,67	5,21
1	46	516	19,75	37,01	28,15	5,02
2	35	224	37,85	60,8	46,47	5,75

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
0	55	837	1,45	16,25	8,79	3,82
1	52	763	17,26	33,05	23,4	4,49
2	30	241	33,88	50,38	41,65	4,51
3	34	169	49,62	84,36	61,66	8,3

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
0	24	271	0,01	17,06	8,83	4,16
1	14	302	18,01	34,94	26,91	4,77
2	7	82	36,96	50,08	43,97	5,16

Figure 8: Obtained results for 8 consecutive periods of time in Guayaquil



Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	52	95	1	10,33	4,99	3,14
2	109	156	10,36	22	16,27	2,84
3	92	110	21,2	32	26,31	3,16
4	53	61	32	46	37,59	3,57
5	24	29	42	58	49,44	3,77
6	21	21	60	113	80,04	3,21

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	34	62	1	10	5,03	2,76
2	107	174	9,2	21,75	16,06	3,26
3	93	135	22	36,5	27,82	3,93
4	49	59	37	54	45,11	4,7
5	12	12	56	113	75,53	3,32

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	47	77	1	11	6,01	3,07
2	106	166	5	22,5	16,53	3,76
3	124	173	21,25	37,67	28,8	4,18
4	48	58	37	55	44,95	5,01
5	13	13	55	70	62,08	4,42
6	8	8	76,5	128	98,53	1,88

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	62	119	1	14	7,65	4,23
2	132	204	14	31,5	20,51	3,71
3	86	112	27,5	44,5	34,83	4,55
4	23	25	45	65	54,03	4,97
5	3	3	65	69	67,33	2,08
6	2	2	79	82	80,5	2,12
7	3	3	88	91,5	89,83	1,76

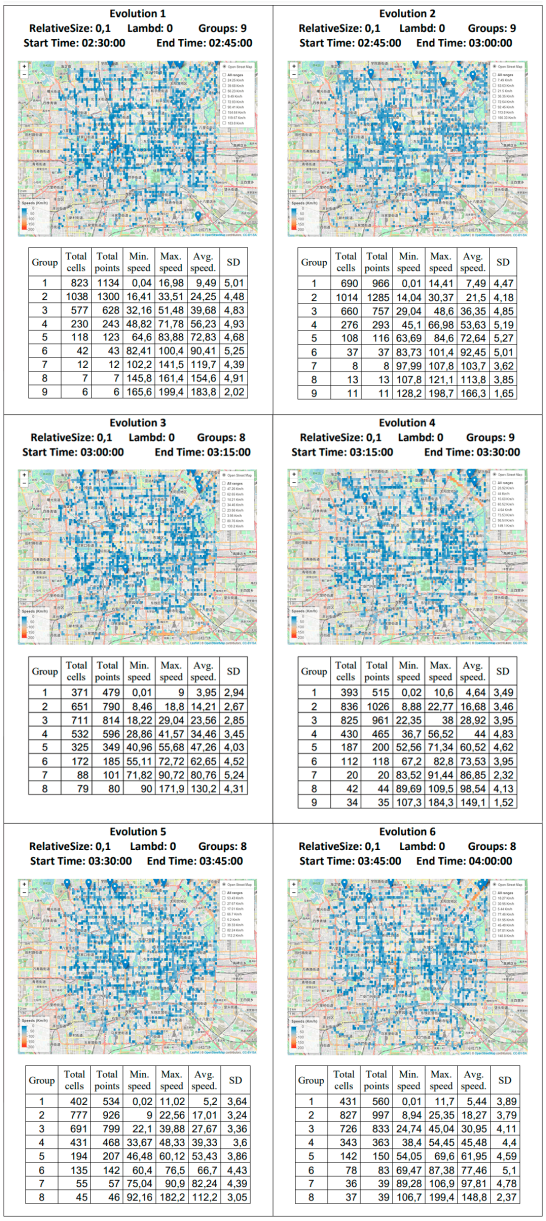
Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	52	90	1	10,17	5,64	3,1
2	112	171	9,1	19,67	15,2	2,65
3	114	142	18,75	34,67	24,36	2,75
4	82	104	28,75	44	34,58	3,97
5	56	64	40,67	119	64,46	3,93

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	110	202	1	16,69	9,97	5,02
2	135	208	16,4	31,5	22,75	4,05
3	92	114	30,5	48	37,64	4,84
4	12	12	46	54,5	50,12	2,88
5	29	32	53	68	60,17	4,55
6	11	11	71,67	126	98,99	1,79

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	90	159	1	15,75	8,85	4,57
2	157	252	15,75	32	23,08	4,52
3	5	5	32	35,5	33,2	1,35
4	72	96	31,67	50,5	39,82	5,51
5	20	22	51	69,5	59,27	5,63
6	14	15	69	128	93,73	2,96

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	55	104	1	12	6,36	3,26
2	128	194	10,17	24,67	17,65	3,16
3	109	149	23	35,6	28,31	3,43
4	73	82	35,5	51,5	42,52	4,61
5	42	45	50,33	129	79,66	3,33

Figure 9: Obtained results for 8 consecutive periods of time in Rome



Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	823	1134	0,04	16,98	9,49	5,01
2	1038	1300	16,41	33,51	24,25	4,48
3	577	628	32,16	51,48	39,68	4,83
4	230	243	48,82	71,78	56,23	4,93
5	118	123	64,6	83,88	72,83	4,68
6	42	43	82,41	100,4	90,41	5,25
7	12	12	102,2	141,5	119,7	4,39
8	7	7	145,8	161,4	154,6	4,91
9	6	6	165,6	199,4	183,8	2,02

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	690	966	0,01	14,41	7,49	4,47
2	1014	1285	14,04	30,37	21,5	4,18
3	660	757	29,04	48,6	36,35	4,85
4	276	293	45,1	66,98	53,63	5,19
5	108	116	63,69	84,6	72,64	5,27
6	37	37	83,73	101,4	92,45	5,01
7	8	8	97,99	107,8	103,7	3,62
8	13	13	107,8	121,1	113,8	3,85
9	11	11	128,2	198,7	166,3	1,65

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	371	479	0,01	9	3,95	2,94
2	651	790	8,46	18,8	14,21	2,67
3	711	814	18,22	29,04	23,56	2,85
4	532	596	28,86	41,57	34,46	3,45
5	325	349	40,96	55,68	47,26	4,03
6	172	185	55,11	72,72	62,65	4,52
7	88	101	71,82	90,72	80,76	5,24
8	79	80	90	171,9	130,2	4,31

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	393	515	0,02	10,6	4,64	3,49
2	836	1026	8,88	22,77	16,68	3,46
3	825	961	22,35	38	28,92	3,95
4	430	465	36,7	56,52	44	4,83
5	187	200	52,56	71,34	60,52	4,62
6	112	118	67,2	82,8	73,53	3,95
7	20	20	83,52	91,44	86,85	2,32
8	42	44	89,69	109,5	98,54	4,13
9	34	35	107,3	184,3	149,1	1,52

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	402	534	0,02	11,02	5,2	3,64
2	777	926	9	22,56	17,01	3,24
3	691	799	22,1	39,88	27,67	3,36
4	431	468	33,67	48,33	39,33	3,6
5	194	207	46,48	60,12	53,43	3,86
6	135	142	60,4	76,5	66,7	4,43
7	55	57	75,04	90,9	82,24	4,39
8	45	46	92,16	182,2	112,2	3,05

Group	Total cells	Total points	Min. speed	Max. speed	Avg. speed.	SD
1	431	560	0,01	11,7	5,44	3,89
2	827	997	8,94	25,35	18,27	3,79
3	726	833	24,74	45,04	30,95	4,11
4	343	363	38,4	54,45	45,48	4,4
5	142	150	54,05	69,6	61,95	4,59
6	78	83	69,47	87,38	77,46	5,1
7	36	39	89,28	106,9	97,81	4,78
8	37	39	106,7	199,4	148,8	2,37

Figure 10: Obtained results for 6 consecutive periods of time in Beijing

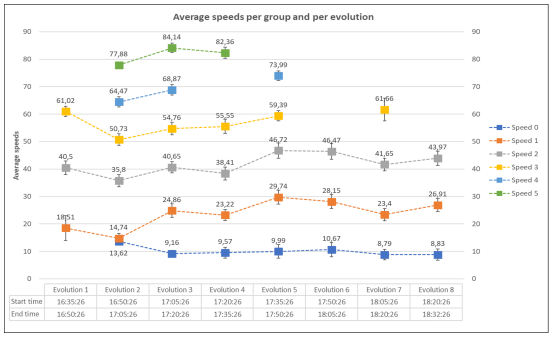


Figure 11: Average speeds by cluster and evolution in Guayaquil

Another aspect to note is the presence of a greater

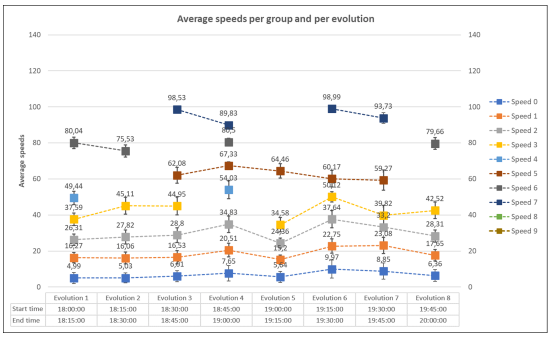


Figure 12: Average speeds by cluster and evolution in Rome

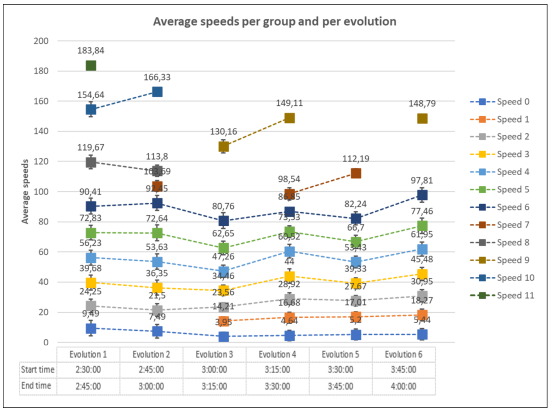


Figure 13: Average speeds by cluster and evolution in Beijing

number of clusters with stable speed ranges in the Beijing dataset (Figure 13) compared to the other two. This is due to the fact that the GPS locations recorded in each trajectory present a more uniform distribution than in the case of Guayaquil and Rome.

The batch processing of the proposed method generates the maps with the information of the cells every n minutes, which for this work was established in blocks of 15 minutes. After that time a map is obtained with the accumulated information within that period of time.

The implementation of the proposed method in a controlled environment can be obtained from a public repository[3].

5 Conclusions

In this article, a methodology has been proposed to dynamically identify the characteristics of the vehicular flow in a period of time. The results of its application for the analysis of vehicular trajectories in the cities of Guayaquil, Rome and Beijing have been satisfactory, and it allowed quickly identifying the different speed

[3]Instructions and source code available at `https://github.com/gary-reyes-zambrano/Dynamic-grouping-of-vehicle-trajectories--DyTRA-`

ranges, mainly those that are formed in lower speeds clusters and with a greater concentration of vehicles.

For this purpose, the trajectory information has been represented in cells and processed using a proposed algorithm based on the methodology defined by the authors of this work [25]. As a result of the clustering process, different clusters were obtained that dynamically change from one evolution to another, identifying common speed ranges at different moments in time.

Interactive maps as part of the methodology are an extremely useful tool when it comes to visualizing the result of the different clusters. They allow displaying specific characteristics of the clusters and analyzing traffic flow in specific sectors. The use of dynamic maps to represent identified speed ranges facilitates the interpretation of vehicular traffic, quickly distinguishing areas with slow traffic that could eventually lead to congestion. In turn, they also quickly bring to the surface atypical situations, generally represented by excessively high speed ranges for what is expected in the region.

Currently, the authors of this work are incorporating into the methodology an improvement with respect to the cell creation process to decrease its processing times.

As lines of future work, we propose the processing the methodology in a parallel architecture to speed up some of the processes such as the visualization of the maps.

It is also proposed to analyze the incremental incorporation of data within the clustering together with the concept of forgetting. In this way, it will be sought to give the clustering the possibility of detecting changes in the behavior of vehicular traffic that will help to identify congestion in a more efficient way.

Competing interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The autor(s) received no financial support for the research, authorship, and/or publication of this article.

Authors' contribution

The authors confirm contribution to the paper as follows: study conception and design: Gary Reyes, Laura Lanzarini, César Estrebou; data collection: Gary Reyes; analysis and interpretation of results: Gary Reyes, Laura Lanzarini, César Estrebou, Aurelio Bariviera; draft manuscript preparation: Gary Reyes, Laura Lanzarini, César Estrebou, Aurelio Bariviera. All authors reviewed the results and approved the final version of the manuscript.

References

[1] A. Jain, "Data clustering: 50 years beyond k-means. 2009," *Pattern Recognition Letters*, 2009.

[2] T. S. Madhulatha, "An overview on clustering methods," *arXiv preprint arXiv:1205.1117*, 2012.

[3] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," 2012.

[4] H. F. Tork, "Spatio-temporal clustering methods classification," in *Doctoral Symposium on Informatics Engineering*, vol. 1, pp. 199–209, Faculdade de Engenharia da Universidade do Porto Porto, Portugal, 2012.

[5] J. Han, M. Kamber, and A. K. Tung, "Spatial clustering methods in data mining," *Geographic data mining and knowledge discovery*, pp. 188–217, 2001.

[6] B. M. Varghese, A. Unnikrishnan, and K. Jacob, "Spatial clustering algorithms-an overview," *Asian Journal of Computer Science and Information Technology*, vol. 3, no. 1, pp. 1–8, 2013.

[7] J. D. Mazimpaka and S. Timpf, "Trajectory data mining: A review of methods and applications," *Journal of Spatial Information Science*, vol. 2016, no. 13, pp. 61–99, 2016.

[8] M. Y. Choong, R. K. Y. Chin, K. B. Yeo, and K. T. K. Teo, "Trajectory pattern mining via clustering based on similarity function for transportation surveillance," *International Journal of Simulation-Systems, Science & Technology*, vol. 17, no. 34, pp. 19–1, 2016.

[9] J. Kim and H. S. Mahmassani, "Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories," *Transportation Research Procedia*, vol. 9, pp. 164–184, 2015.

[10] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data - SIGMOD '07*, p. 593, ACM Press, 2007.

[11] Y. Mao, H. Zhong, H. Qi, P. Ping, and X. Li, "An adaptive trajectory clustering method based on grid and density in mobile pattern analysis," *Sensors*, vol. 17, p. 2013, Sep 2017.

[12] G. Yuan, P. Sun, J. Zhao, D. Li, and C. Wang, "A review of moving object trajectory clustering algorithms," *Artificial Intelligence Review*, vol. 47, p. 123–144, Jan 2017.

[13] L. X. Liu, J. T. Song, B. Guan, Z. X. Wu, and K. J. He, "Tra-dbscan: A algorithm of clustering trajectories," in *Frontiers of Manufacturing and Design Science II*, vol. 121 of *Applied Mechanics and Materials*, pp. 4875–4879, Trans Tech Publications Ltd, 1 2012.

[14] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.

[15] G. Reyes-Zambrano, L. Lanzarini, W. Hasperué, and A. F. Bariviera, "GPS trajectory clustering method for decision making on intelligent transportation systems," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 5529–5535, 2020.

[16] Q. Yu, Y. Luo, C. Chen, and S. Chen, "Trajectory similarity clustering based on multi-feature distance measurement," *Applied Intelligence*, pp. 2315–2338, 2019.

[17] T. Luo, X. Zheng, G. Xu, K. Fu, and W. Ren, "An improved dbscan algorithm to detect stops in individual trajectories," *ISPRS International Journal of Geo-Information*, vol. 6, no. 3, 2017.

[18] N. Ferreira, J. T. Klosowski, C. Scheidegger, and C. Silva, "Vector field k-means: Clustering trajectories by fitting multiple vector fields," 2012.

[19] H. Hu, G. Lee, J. H. Kim, and H. Shin, "Estimating Micro-Level On-Road Vehicle Emissions Using the K-Means Clustering Method with GPS Big Data," *Electronics*, 2020.

[20] J. Lou and A. Cheng, "Detecting pattern changes in individual travel behavior from vehicle gps/gnss data," *Sensors*, vol. 20, no. 8, 2020.

[21] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, (New York, NY, USA), p. 1–16, Association for Computing Machinery, 2002.

[22] M. Garofalakis, J. Gehrke, and R. Rastogi, *Data Stream Management: Processing High-Speed Data Streams*. Data-Centric Systems and Applications, Berlin, Heidelberg: Springer Berlin Heidelberg, 2016.

[23] M. R. Ackermann, C. Lammersen, C. Sohler, K. Swierkot, and C. Raupach, "StreamKM++: A clustering Algorithm for Data Streams," *ACM Journal of Experimental Algorithmics*, vol. 17, pp. 173–187, 2012.

[24] C. C. Aggarwal, *Data Streams: An Overview and Scientific Applications*, pp. 377–397. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[25] G. Reyes, L. Lanzarini, C. Estrebou, and V. Maquilón, *CACIC 2021 XXVII CONGRESO ARGENTINO DE CIENCIAS DE LA COMPUTACION*, vol. XXVII, p. 261–270. Universidad Nacional de Salta, 2021 ed.

Citation: G. Reyes, L. Lanzarini, C. Estrebou and A. Bariviera. *Dynamic grouping of vehicle trajectories.* Journal of Computer Science & Technology, vol. 22, no. 2, pp. 141–150, 2022.
DOI: 10.24215/16666038.22.e11
Received: April 4, 2022 **Accepted:** August 29, 2022.
Copyright: This article is distributed under the terms of the Creative Commons License CC-BY-NC.