07: Detecting a traffic jam using trajectory data

1st Sijie Zeng
CEG Faculty, TU Delft
Delft, Netherlands

2nd Kirsten Bos
CEG Faculty, TU Delft
Delft, Netherlands

***Abstract*—In this paper, the purpose is to capture the traffic patterns from the pNEUMA open-source trajectory data in Athens. We want to provide a basic threshold which indicates a traffic jam. This can be done by visualizing the area of interest using k-means clustering. Before clustering can be applied, the data is preprocessed to extract important features from the data, which are *speed*, *density* and a combination of *speed* and *density*. K-means clustering will create cluster labels which can be replaced with the mean value of the cluster. This value can be used to analyze the traffic patterns. Analyzing the plots of the traffic patterns, a threshold value is obtained that can describe a traffic jam in Athens. The threshold value for *speed* is a speed below *4 m/s*, and the threshold value for *density* is between *5 and 10 vehicles passing the grid*. If both conditions are reached, the traffic pattern can be considered as a traffic jam. It was not possible to come up with a threshold value for the combination of *speed* and *density*.**

***Index Terms*—data driven, k-means clustering, traffic patterns, traffic jam, trajectory data, PCA**

I. Introduction

With more and more traffic in the metropolitan nowadays, the amount of urban traffic jams is increasing. These traffic jams are wasting time and energy, causing higher CO2 emissions and road rage, decreasing productivity, and imposing costs on society [1]. Most of previous research is focused on traffic jams on highways because usually the urban traffic jam has its own exclusivity and regional specificity. For example, the road network (number of lanes, number of intersections, traffic direction) and the composition of traffic (pedestrians, bicycles, vehicles) is much more complex than the highway, and it also differs from city to city. Thus, we cannot give a general definition of traffic jam to all the cities. Therefore, in this research, the trajectory data of Athens is studied for finding traffic patterns and try to give a definition of a traffic jam in Athens. The area of interest is shown in figure 1. The main goal is to detect a traffic jam. Different features can be used to determine a traffic jam, but this research focus on *speed*, *acceleration*, *density* and *direction*.

Detecting a traffic jam using trajectory data is useful for society. If a traffic jam occurs frequently on a certain location, it might be necessary to make some adjustments on the road infrastructure. This will reduce the negative influence of the traffic jam on humans. However, not all road users will be happy if their trajectory data is used, because it can harm the privacy of people if their license plate is published. This sensitive information is already filtered out before the dataset is published. As stated before, it is very important to keep in mind the used method in this research might not be suitable for all urban areas.

To address the issue of how to identify the urban traffic jam from trajectory data, the unsupervised method, k-means clustering, is used in this research due to the use of unlabelled trajectory data.

The rest of this paper is organised as follows. In section II, a flowchart of the methodology as well as its justification are presented. Section III discusses the results of the clustering and the patterns analysis. Section IV justifies the method selection, analyses the results to detect the traffic jam and concludes the threshold value of a jam while section V draws a conclusion of the project and points out some potential future research.

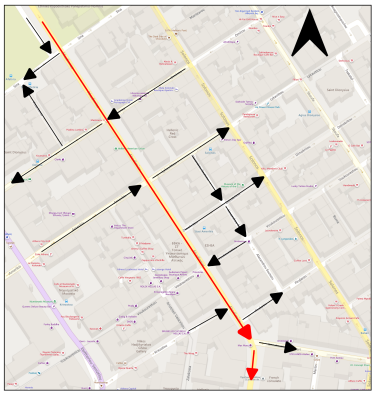

Fig. 1. The area of interest in Athens including the heading direction of the vehicles (created using QGIS3). The red arrow indicates the main road, and the black arrows indicate the side roads

II. Methods

This section is going to present the data description, data preprocessing, methodology and the justification of the approach. The flowchart of the research is shown in figure 2.

A. Data description

The used dataset is extracted from pNEUMA, which contains multi-modal trajectory data of a downtown area of Athens, Greece. There are different datasets available, but this research only uses the data of region 1 of Nov 1, 8:00-8:30 am [2]. The data is obtained using drones, which hover over the area of interest [3]. The drones measure different kind of vehicles, like cars, buses and powered two wheelers. The dataset is preprocessed and in this preprocessing the geographical coordinates are transformed to flatten plane coordinates. Thereby,

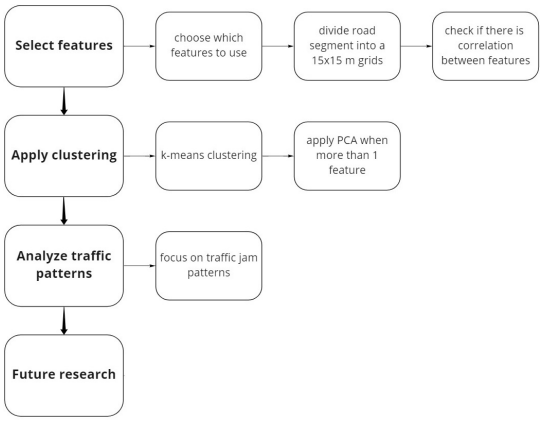

Fig. 2. Flowchart of the research (created using miro).

the heading directions are computed for each moment. The resulting dataset is displayed in figure 3.

	track_id	frame_id	lat	lon	speed	lon_acc	lat_acc	time	E	N	E_heading	N_heading
0	1	960	37.977	23.737	3.744	0.047	-0.066	38.40	740406.453	4206851.876	0.693	0.354
1	1	961	37.977	23.737	3.745	0.038	-0.071	38.44	740406.541	4206851.879	0.693	0.354
2	1	962	37.977	23.737	3.745	0.033	-0.076	38.48	740406.714	4206851.995	0.693	0.354
3	1	963	37.977	23.737	3.745	0.030	-0.081	38.52	740406.801	4206851.997	0.693	0.354
4	1	964	37.977	23.737	3.746	0.025	-0.086	38.56	740406.974	4206852.113	0.693	0.354

Fig. 3. Table containing the first 5 rows of the dataset which is used (created using python).

B. Data prepocessing

The preprocessing of the data consists of two steps: determining the representative features, transforming them into a feasible and reasonable form and calculating the correlation.

1) Features: Based on the aim of detecting a traffic jam, four features are extracted from the dataset, which are *speed*, *density*, *acceleration* and *direction (confusion)*. The original dataset is showing the trajectory of each individual vehicle, however, a traffic jam or congestion is a traffic condition at the macro level which requires aggregation of vehicles. Therefore, we determine to divide the road network into 15m * 15m grids (see figure 4) and aggregate the data for each grid. Furthermore, considering the space-time characteristic of traffic jam, the data of the four features should also be transformed into the form of time-space.

Take *speed* as an example, we calculate the average speed of the vehicles for each grid and all the time intervals. The NaN value indicates that there is no vehicle and will be padded with a value of free speed which is 30 m/s, because this is approximately two times the maximum speed which occurs, which is 16 m/s. We assume a value twice times bigger than the maximum is large enough to divide the NaN values from the real values.

Just like *speed* using mean value for each grid, we need to determine the feature value for other three attributes. *Density* is not the theoretical density, instead, the value is the number of vehicles passing by the grid during the time interval. The NaN will be padded with zero which indicates no vehicle is passing.

The *acceleration* can be calculated using the speed (v [m/s]) and the time (t [s]), see equation 1. The acceleration is calculated for each vehicle, but the first and the last value for each vehicle are set to NaN, because these calculations use values of another vehicle. The mean acceleration is calculated for each grid cell. The NaN values are replaced with a value of 10 m/s^2, because the maximum acceleration which occurs is around 5 m/s^2 (so two times the maximum acceleration). We assume a value twice times bigger than the maximum is large enough to divide the NaN values from the real values.

$$a = \frac{\Delta v}{\Delta t}[m/s^2] \tag{1}$$

The *direction (confusion)* is calculated using *E heading* and *N heading* which is Cartesian coordinates in the original dataset and converted to polar coordinates, see equation 2. We are interested in the angle (ϕ), which is converted from radians into degrees. If a lot of vehicles are turning direction in one grid, the area has a higher degree of confusion and may be an intersection. An intersection may have a higher probability to have a traffic jam, thus the standard deviation of the *direction (confusion)* is calculated for each grid cell. The NaN values are replaced with a value of zero because if there is no vehicle passing, the grid will have the lowest degree of confusion.

$$\rho = \sqrt{(E_{heading})^2 + (N_{heading})^2}$$

$$\phi = arctan(\frac{N_{heading}}{E_{heading}}) \tag{2}$$

The grid values of the four different features are each stored in different datasets, containing all the 26 time intervals which are each 28.8 seconds.

For follow-up, all the different features are merged for a specific time interval. This results in 26 datasets containing the *E id*, *N id* (location of the grid cell), the *speed*, the *density*, the *acceleration* and the *direction (confusion)*.

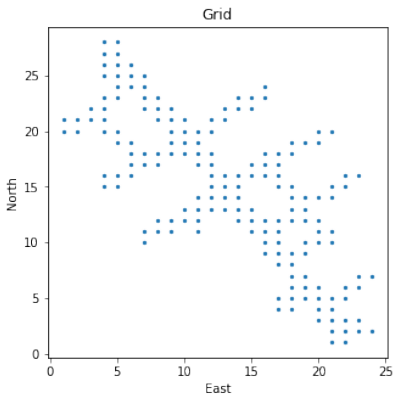

Fig. 4. Example of the created grid (created using python).

2) Correlation: To capture as many features as possible, the data will be clustered using different combinations of features. As discussed before, the chosen features are *speed*, *density*,

acceleration and *direction (confusion)*. Before applying clustering, the correlation between the features is considered. If the values are highly correlated, so a correlation around 1 or -1 [4], it is better to remove one of the features, because using similar features would give a similar result. Therefore, feature selection is applied, which removes irrelevant (in this case similar) features so there will be fewer combinations which speeds up the process and reduces the maximum number of dimensions [5]. So, the feature *speed* and *acceleration* will not be used in the same feature combination.

C. Classification using k-means clustering

The data does not contain labels, so the machine learning method will be unsupervised. According to [6], the k-means clustering is among the most adaptable clustering techniques for railway transportation data. In [7], the hierarchical clustering is used to process taxi GPS data. Among all the different unsupervised methods, hierarchical clustering and k-means clustering are considered to classify the data. After comparing the two methods discussed in section IV, there is decided to only apply k-means clustering to classify the data. After the clustering, the patterns can be analysed.

For the combination of features, there are 4 possible dimensions

- 1-dimensional: only use 1 feature
- 2-dimensional: use 2 features
- 3-dimensional: use 3 features
- 4-dimensional: use all the 4 features

Which combinations we use depends on the result of the correlation which was discussed in the preprocessing.

When applying k-means clustering, each point in the grid gets a label which indicates the number of the cluster. The labels vary from *0* to *k-1*, where *k* is the number of clusters. We cannot determine traffic patterns using labels which are assigned randomly for each time interval. For example, a label of 1 might indicate NaN values for time interval 1, but for time interval 2 a label of 1 might indicate a speed of 5 m/s. Therefore, the by the algorithm generated labels are replaced with the mean cluster value, as shown in figure 5. For example, for feature *speed*, cluster 1 consists of 3 grid cells with a speed of 4 m/s, 5 m/s and 6 m/s respectively, then the cluster value will become 5 m/s instead of the number of the cluster. So, the mean cluster value of each cluster is calculated, and this value is added to the dataframe. The mean cluster value is then used as the color for the plot instead of the cluster label. If we use these meaningful values for visualization, it will be possible to search for traffic patterns in the data, like a traffic jam.

For 2-dimensional (and higher orders) clustering, the PCA method is applied to combine the features and obtain a cluster label. As stated before, this value is quite random and cannot be used to define the traffic patterns. However, it is not possible to just take the mean value of the feature. In the case of more than 1 dimension, there are multiple features in which each feature has different units, so they cannot be combined easily.

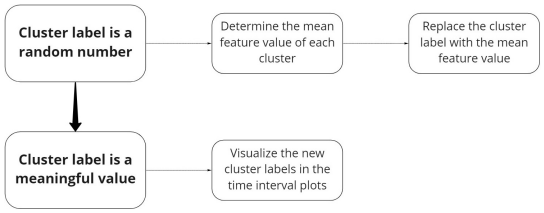

Fig. 5. Convert the random cluster labels to meaningful labels (created using miro).

Therefore, the following steps are performed to get a mean value

- Set the NaN values to *NaN*, so a *speed* of *30 m/s* and an *acceleration* of *10 m/s*^{2} will become *NaN*
- Normalize the data from the features using min-max normalization, so all features will get values varying from 0 to 1
- A green value (0) indicates good flow, and a red value (1) indicates bad flow, so the possibility of a traffic jam. *Speed* and *acceleration* must be transformed, because for these features a low value (0) indicates bad flow and a high value (1) indicates good flow. This can be done using the following formula

$$1 - \textit{normalized feature value}$$

- Take the mean value of the normalized features, so there is only 1 value for each label: average of the features
- Replace the cluster label with the mean of the average of the features

T o analyse the traffic patterns, timeseries of the k-means clustering are created. For simplicity, the timeseries consist of 9 time intervals instead of 26.

After the classification, the patterns are analysed and there is detected which clusters can be defined as a traffic jam and which are not. Possible indications of a traffic jam are

- Low speed
- A negative or a very low positive acceleration
- High density

III. Results

This section is going to present the obtained results. First, the results of the correlation calculation are shown, which leads to the possible feature combinations. The feature combinations can be used to apply the k-means clustering.

A. Correlation

The correlation between the different features are calculated using the *.corr(method='pearson')* function in Python, which calculates the standard correlation coefficient. The results are similar for the different time intervals. From the table in figure 6, it can be seen there is correlation between the features when the cell has a red color. So, there is a high correlation between the *speed* and *acceleration*, and the correlation between other features is quite low.

The *direction* will not be considered for feature combination. The reason why we do not take *direction* into consideration is because the *speed* and *density* are more useful for detecting a traffic jam, and the *direction* is showing the confusion degree of the grids. Study the clustering of standard deviation is more useful when detecting the intersection and study the traffic patterns at the road area that needs to turn direction.

After looking at the correlation, the following features and feature combinations are considered: *speed*, *density*, *direction*, and the combination of *speed* and *density*. So, for 1-dimensional there are 3 possible features, for 2-dimensional there are 1 possible feature combinations and we do not have 3-dimensional or 4-dimensional feature combination.

	Feature1_speed	Feature2_density	Feature3_acceleration	Feature4_direction
Feature1_speed	1.000000	-0.568711	0.981633	-0.444584
Feature2_density	-0.568711	1.000000	-0.604375	0.312769
Feature3_acceleration	0.981633	-0.604375	1.000000	-0.375502
Feature4_direction	-0.444584	0.312769	-0.375502	1.000000

Fig. 6. The correlation between the different features for time interval 0 (created using python). The correlation matrices for the other time intervals are similar to this one.

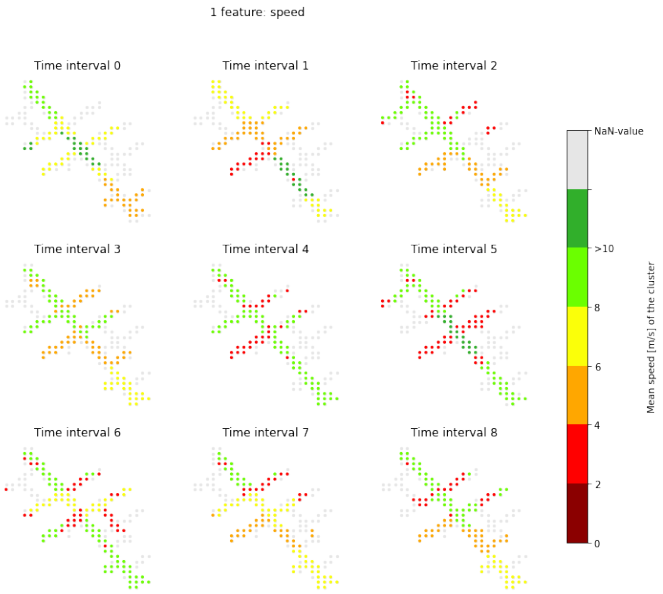

Fig. 7. K-means for 9 time intervals using feature speed (created using python).

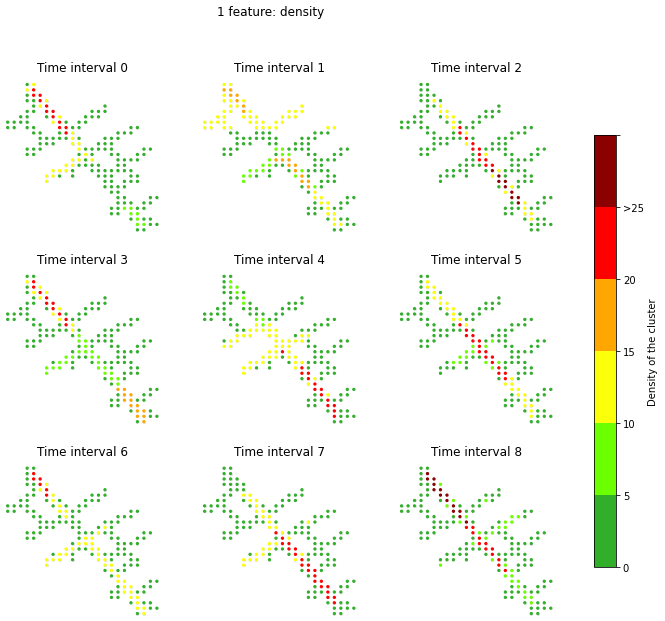

Fig. 8. K-means for 9 time intervals using feature density (created using python).

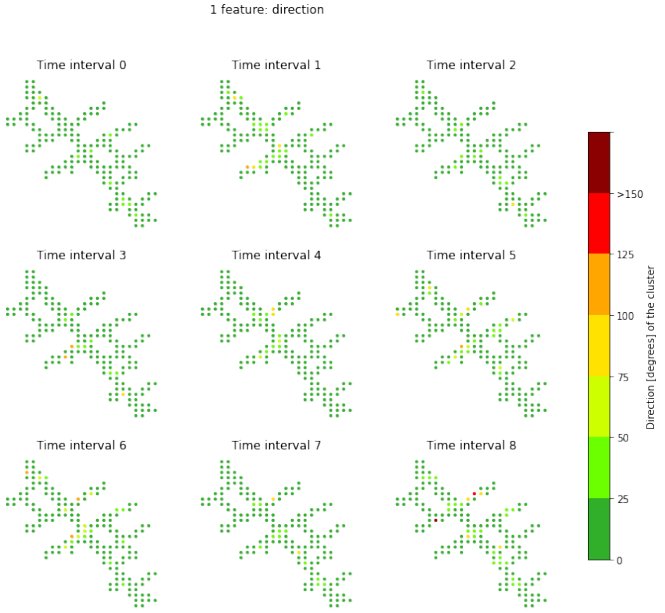

Fig. 9. K-means for 9 time intervals using feature direction (created using python).

B. K-means clustering

K-means clustering is applied to all grids at each time interval, using a k of 8 (so the number of clusters will be 8). This gives 8 different cluster labels, which are replaced by the mean value of the cluster. The result of the clustering is visualized in figure 7 to 10. The amount of colors in these figures might be less than the number of clusters, because the label can be replaced by similar values. For example, feature *speed*, there might be a mean cluster value of *1 m/s* and a value of *1.5 m/s* which will both be displayed with a dark red color. However, this does not matter because we are interested in all the low values because this might indicate a traffic jam.

Figure 7 shows the results of one dimension feature of *speed* from time interval 0 to 8. The legend on the right represents the color of each cluster as well as the mean value of each cluster. The grey color indicates a NaN value, so there is no vehicle present in the grid point. Figure 8 and figure 9 show the result of one dimension feature of *density* and *direction*, while

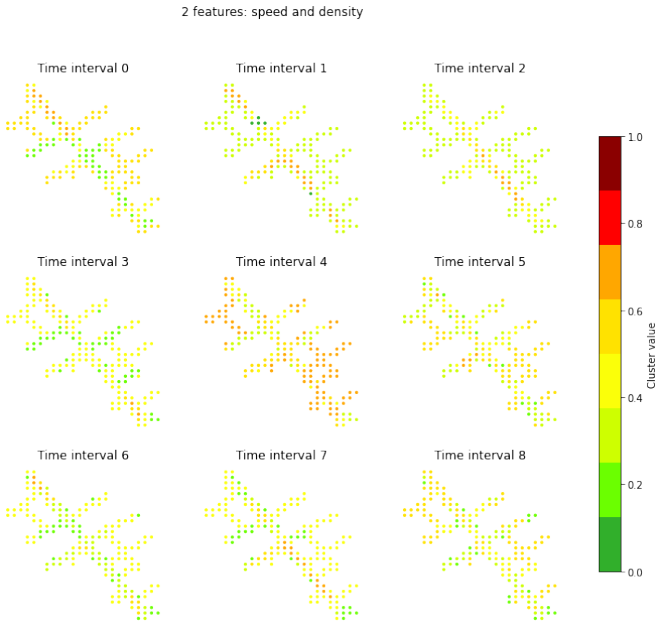

Fig. 10. K-means for 9 time intervals using feature density and speed (created using python).

figure 10 shows the two dimensional feature combination of *speed* and *density*.

C. Analysing the traffic patterns

As shown in figure 4, the 2D plot shows the shape of the study area. The driving direction of the vehicles is shown in figure 1. The road network consists of a main road (from northwest to southeast) and several intersections (e.g., the first one is evacuating from main road to southwest; the second one is from north east merging into the main road and then evacuating to southwest). The features *speed*, *density* and *direction* are analysed and also the feature combination of *speed* and *density* to analyse the traffic patterns.

Speed: First appears red grids on time interval 1 on the branch of the third intersection where the vehicles are merging into the main road, in the next time interval, the vehicles are evacuating. From time interval 4 to 8, the vehicles coming from the third intersection southwest are merging into the main road, and then the flow is going to the northeast from the third intersection. The vehicles from the northeast branch of the second intersection are congested and merge into the main road and go to the southwest branch of the second intersection. In time interval 6, the vehicles from the northeast branch of the third intersection are evacuating to a southeast branch and cause congestion. In time interval 7, the branches are evacuating, but the northeast branch of the second intersection still has vehicles coming in and propagates to the southwest branch of the second intersection in time interval 8.

Density: The maximum value of density shown in time interval 0 to 8 is above 25 vehicles per grid. As shown in time interval 8, the density in the upstream of the main road is large, but the speed is not low in the same area at the same time, which indicates that the speed is more representative for a congestion than density. In general, the branch roads do not have a high density, which indicates it will not have lots of vehicles in the grids of branch road, at the same time, the speed could have a low value, which indicates in real life, the vehicles in branch roads will have a slower speed.

Direction: In general, the direction plots show that at intersections, the confusion degree is higher than in other areas, if it is the same, then it might be a red light at the signalized intersection, or no vehicles are passing.

2D-Speed and density: Based on the calculation of applying PCA, the red point indicates the congestion. In time interval 0, from the upstream of the main road is orange, which indicates a higher probability of congestion. However, the speed is relatively high and the density is relatively high in this area, thus, it could not give a clear guideline for defining the traffic jam.

IV. Discussion

In this section the choice of the clustering method is discussed. Thereby an explanation is given for the number of clusters and why certain features are chosen for visualization. Looking at the visualization of k-means clustering, traffic patterns can be recognized.

A. K-mean clustering or hierarchical clustering?

The feature speed and a time interval from 0 to 5 are chosen to compare the different clustering methods. When comparing k-means with hierarchical clustering it can be seen there is a difference between the two methods. To determine which method is better for the purpose of determining traffic patterns, the clustered values are compared to the actual values. Looking at time interval 5 in figure 11, the green part in the middle is also visible in the k-means clustering. From this, it can be concluded k-means clusters the feature better than hierarchical clustering.

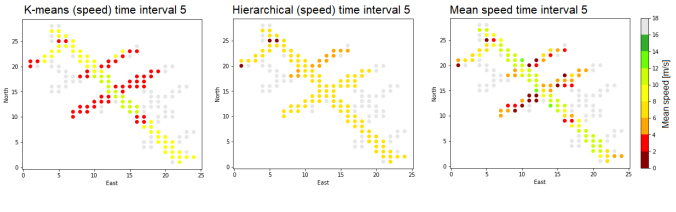

Fig. 11. Comparison between hierarchical clustering, k-means clustering and the actual mean speed of the grid cell (created using python).

B. Number of clusters

In this research, the number of clusters is set to be 8 for all features and feature combinations. This is done for comparison between the features. However, it might not be the best approach. It could be using the optimal number of clusters for each feature and each time interval gives a better result. However, we are not really interested in the best way of clustering, we are more interested in the traffic patterns in the data. The impact of a difference in the number of clusters could be a topic for further research.

C. Traffic patterns

Speed: From the subsection of analysing the traffic patterns in section III, the cluster of red grids which contains a mean speed value of 0 to 4 m/s is the cluster that we have most interests in, shows a congested traffic condition. Therefore, a threshold value of 4 m/s can be an upper limit of a traffic jam in Athens.

Density: For *density*, the value represents the number of vehicles passing by the grids, several declarations should be clarified: 1) In Google Maps, most of the branch roads are having single road lane, and the width of a road lane is usually around 3 and 5 meters, which may affect the validation of the data used for clustering. 2) The road capacity is different for different road conditions, but in this research, it is assumed to be the same. If the *density* is low, it could be a jam in which only a few vehicles can pass, or it could be a free flow condition where only a couple of vehicles are passing. Thus, from a comprehensive consideration it can be concluded that if a low speed and low density happen at the same time, it could be defined as a traffic jam. The traffic condition of high density and low speed can also be considered, while the condition of high density and high speed should be the free flow condition. Based on the analysis of *speed*, in time interval 1, the southwest branch road of the third intersection has both low speed and low density which can be considered as traffic jam. In time interval 4, the speed of the southwest branch road of the third intersection is low, but the density is around 15 vehicles per grid, which can not be defined as a jam. In time interval 5, the northeast branch road is definitely a jam. In conclusion, if the speed is lower than 4 m/s and the density is lower than 10 vehicles per grid and higher than 5 vehicles per grid, it could be considered as a traffic jam.

2D-Speed and density: In the 2D-speed and density plot, we could not see a clear consistency of the results along with the 1D-speed and 2D-density. Therefore, we would not give a threshold value for the definition of a traffic jam.

V. Conclusion

This research is focusing on detecting a traffic jam using clustering and processing the trajectory data to find the location of a traffic jam. In this case, both k-means and hierarchical clustering are potential methods [7] [6] to classify the data. After comparison (see section IV), it can be concluded that k-means is more suitable than hierarchical clustering for determining a traffic jam, because the clusters generated by k-means look more like the actual data than the clusters of hierarchical clustering.

K-means clustering is then applied on a grid of the area of interest which is Athens. The data can be obtained from pNEUMA. There are different datasets available (10 locations, 4 dates, and 5 timeslots of 30 minutes), but in this research we only use the first 4 minutes of the data of location 1 of Nov 1, 8:00-8:30am. The k-means clustering method gives us labels, to which we can assign the mean value of the cluster. For each grid we will analyse 9 intervals of 28.8 seconds. Possible future work is up scaling the area of interest, so look at different days, longer time intervals and a larger location.

For clustering different (number of) features can be used. Initially, 4 different features are observed (*speed, density, acceleration, direction*). The features can be used separately, but multiple features can also be combined. Looking at the correlation between the features, the combination using both *speed* and *acceleration* could be removed due to high dependency. This research visualizes time interval plots of the *speed*, *density* and the combination of the *speed* and *density*. The time intervals can then be used for analysing the traffic patterns. The *direction* is used to find the locations of intersections.

In this research, the 3 dimensional combination is not applied. The reason is that when we focus on detecting a traffic jam, the *speed* and *density* are the most common features for analysis, thus we consider not only each of them but also a 2-dimensional combination of both. However, the *direction(confusion)* is analysed for an interest of pointing out where is at high degree of confusion. Therefore, the 1-dimensional clustering of *direction(confusion)* is applied, but the combination of *speed* with *direction(confusion)*, *density* with *direction(confusion)* and the 3-dimensional combination are not analysed.

In terms of the threshold value of a traffic jam in Athens, we conclude that when *speed* lower than 4 m/s as well as the *density* is around 5 to 010 vehicles passing a grid can be defined as a traffic jam. For the combination of the features *speed* and *density*, the result is not as expected. The potential reason might be the logic of prepocessing and applying PCA is not suitable for capturing the traffic jam patterns when combining multiple features. Furthermore, it is much harder to give a meaningful value to the clusters when using multiple features (especially when using more than 2 features). It might be a good idea to look into a better method how to obtain one value when using multiple features.

References

[1] Robert L Bertini. You are the traffic jam: an examination of congestion measures. In *The 85th annual meeting of transportation research board*. Citeseer, 2006.

[2] E. Barmpounakis and N. Geroliminis. Pneuma dataset. https://open-traffic.epfl.ch/index.php/about/. Accessed 2020.

[3] Emmanouil Barmpounakis and Nikolas Geroliminis. On the new era of urban traffic monitoring with massive drone data: The pneuma large-scale field experiment. *Transportation research part C: emerging technologies*, 111:50–71, 2020.

[4] Deborah J Rumsey. How to interpret a correlation coefficient r. *Statistics For Dummies*, 2016.

[5] Husna Aydadenta and Adiwijaya Adiwijaya. A clustering approach for feature selection in microarray data classification using random forest. *Journal of Information Processing Systems*, 14(5):1167–1175, 2018.

[6] Faeze Ghofrani, Qing He, Rob MP Goverde, and Xiang Liu. Recent applications of big data analytics in railway transportation systems: A survey. *Transportation Research Part C: Emerging Technologies*, 90:226–246, 2018.

[7] Yulong Wang, Kun Qin, Yixiang Chen, and Pengxiang Zhao. Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi gps data. *ISPRS International Journal of Geo-Information*, 7(1):25, 2018.