2014 13th International Conference on Control, Automation, Robotics & Vision
Marina Bay Sands, Singapore, 10-12th December 2014 (ICARCV 2014)

Time-dependent Partitioning of Urban Traffic Network into Homogeneous Regions

Antonis F. Lentzakis*, Rong Su*, and Changyun Wen*
*School of Electrical and Electronic Engineering,
Nanyang Technological University,
Singapore
Corresponding author: antonios001@e.ntu.edu.sg

***Abstract*—Congestion in urban areas constitutes an important problem that affects people in explicit but also implicit ways. Current research literature on Urban Traffic Estimation has shown that homogeneous distribution of vehicle density along the links of urban traffic networks plays an important role in the derivation or even the existence of the so-called Urban-Scale Macroscopic Fundamental Diagram or MFD in short. This Urban-Scale MFD can provide information that facilitates the application of perimeter traffic control strategies. In this paper, we implement a partitioning of an urban road network into homogeneous regions based on historical traffic information. Using prior information, we make informed decisions about the selection of the region on which the urban road network is based on, as well as the particular time periods for which the partitioning is to be implemented. We make use of weighted *k*-means, *k*-harmonic means and normalized spectral clustering techniques to successfully partition the region into clusters defined by low link density variability, while ensuring that the resulting partitions are spatially cohesive.**

***Keywords*—Congestion, Urban-scale MFD, *k*-means, *k*-harmonic means, spectral clustering, silhouette coefficient, historical data, link densities**

I. Introduction

Congestion in urban road networks is a complex problem that is inextricably linked with quality of life. The socio-economic impact of congestion is so great, that the European Commission stated in a White Paper [1] that the annual costs associated with congestion will accumulate to €80 billion in the European Community alone. Due to the high cost of developing new infrastructure, optimal utilization of available infrastructure is required. The dynamics of traffic flow and congestion formation have been modeled at different scales according to formulations based on fluid mechanics, cellular automata, particle systems, queuing systems, among others. The majority of these models constitute an approximation of the traffic flow behavior, because the unpredictability of human behavior poses considerable limitations to the exact reproducibility of the actual system behavior. Godfrey [2] first proposed the existence of a Macroscopic Fundamental Diagram representing the relationship between vehicle density (veh/km) and space-mean flow (veh/h) for urban regions. Herman and Prigogine [3] defined a Two Fluid Kinetic model later improved upon by Herman and Ardekani [4] that stated that the average speed in urban network traffic is a function of

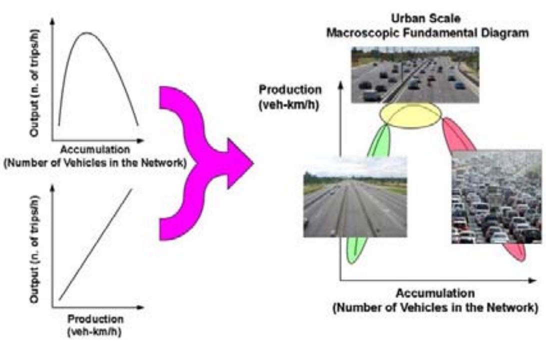

Figure 1. Illustration of the postulates set by Geroliminis and Daganzo [6] based on the experimental results from Yokohama and the microscopic simulation results run on a network representing Downtown San Francisco

the ratio of stopped vehicles, which can be represented as a power function of network link density. Daganzo [5] defined a relationship between urban network output, representing the number of trips ending inside or outside of the network, and number of vehicles present at the network, or accumulation, for brevity. This relationship should be valid under the condition that congestion is distributed in a homogeneous fashion throughout the network and as long as external conditions are varying at a slow rate. Geroliminis and Daganzo [6] further extended this by postulating that a Macroscopic Fundamental Diagram exists for urban regions characterized by homogeneously distributed congestion, the urban network output is proportional to the product of average flow and network length, or production, for brevity. See Fig. 1 for an illustration. Geroliminis and Daganzo [7] employed microscopic simulation of the San Francisco downtown area during peak period for different demand distribution scenarios, as well as analyzing experimental data derived from a 10 km^2 area in Yokohama, Japan and were able to posit the existence of an Urban-scale Macroscopic Fundamental Diagram independent of traffic demand. Geroliminis and Sun [8], Mazloumian et al. [9], Daganzo et al. [10] have shown that variability of network link density plays a major role in the potential shape, scatter or existence of the Macroscopic Fundamental Diagram.

Therein lies the motivation of partitioning urban traffic networks into homogeneous regions. It is well documented that different traffic patterns appear throughout the day. It follows that different partitions of the same urban traffic network need to be attributed to different time periods. Implementation of partitioning can be based on historical data, real-time data, different demand scenario-based micro-simulations or any combination thereof. Ji and Geroliminis [11] chose the scenario- based micro-simulation approach for a 6.5 km^2 area of San Francisco consisting of about 100 intersections and around 400 network links. Their choice enabled them to pinpoint appropriate times when network congestion has not set yet and the traffic volumes are still high, providing them with a variety of link densities for their algorithms. Our network, representing the Marina Bay area of the city of Singapore, also incorporates parts of Chinatown, Clarke Quay, Dhoby Ghaut and the Central Business District. In this paper, we elected to use historical data from loop detectors situated upstream of the majority of intersections for all directions. While historical data can provide more accuracy than scenario-based simulations, as regards the link density values, the question arises as to how to pinpoint the appropriate times with conditions that guarantee adequate link density variability to warrant partitioning of the selected network. Aslam et al. [12] used loop detector data obtained from the Land Transport Authority in Singapore and collected travel data from taxis for the month of August 2010. Machine learning was employed for vehicle distribution and traffic volume inference for the entire island of Singapore. Thus, they were able to identify hotspot areas throughout Singapore for specific time periods. Their results enabled us to select times for our network which can guarantee the link density variability required for successful partitioning of our network into homogeneous regions. In this paper, we derive a partitioning for our network using weighted *k*-means clustering, *k*-harmonic means clustering and spectral clustering algorithms. We utilize a performance metric developed by Ji and Geroliminis [11] to evaluate the intra- and inter-cluster link density variability and we also test for the spatial cohesion of each resulting cluster. The results show that the optimal number of clusters for this particular network is 2, while *k*-harmonic means clustering and normalized spectral clustering provide the most stable results.

II. ALGORITHM DESCRIPTION

A. Center-Based Clustering using k-means, k-harmonic means

The *k*-means algorithm belongs to the category of center-based clustering algorithms [13]. Center-based clustering algorithms are usually characterized by their objective function. Each algorithm tries to minimize the objective function iteratively, updating the cluster centers in every iteration, until a local minimum is observed. The *k*-means algorithm clusters data points in *k*-sets, where *k* represents the number of cluster centers. Voronoi partitions are formed according to data point membership to each cluster. The *k*-means algorithm uses the Euclidean norm to compute the distance of each data point from the selected cluster centers and then updates the centers according to the means of the data points associated with that cluster. Weighted *k*-means uses feature weights to signify the importance of one particular feature over others. In our implementation, we define w_c as the weight for the link coordinates feature and w_ρ as the weight for the density feature. Thus we have:

$$WKM(\mathrm{D},\mathrm{M}) = \sum_{i=1}^{n} \sum_{f\in\{c,\rho\}} \min_{j=\{1,\ldots,k\}} w_f \left\| d_{if} - m_{jf} \right\|^2 \tag{1}$$

where we have $d \in D \in \mathrm{R}^{n\times 3}$ representing the data points, $m \in M \in \mathrm{R}^{k\times 3}$ representing the cluster centers and k is the number of clusters. The reason we enforce weights is to ensure the spatial cohesion of the resulting partitions, otherwise, the resulting clusters will consist of spatially disparate links with admittedly very low density variability. One other main caveat in employing *k*-means clustering lies in the fact that the resulting data point clusters may not represent global minima, as regards to the feature variability we are trying to minimize and are highly dependent on the initialization of the cluster centers. This weakness can be partially mitigated if we select, using prior knowledge, particular data points as the initial cluster centers, but information at this level of detail is usually hard to obtain.

We chose to apply a weighted variation of *k*-harmonic means algorithm as the resulting clusters are insensitive to cluster center initialization. The *k*-harmonic means algorithm [14] objective function is characterized by the use of the harmonic mean instead of the Euclidean norm of the distance of each data point from the cluster centers. Although *k*-harmonic means includes a weight function that rewards data points with greater distance from their respective cluster centers, due to the particular nature of partitioning we aim to implement, we include feature weights as in the case of the k-means algorithm. Thus we have:

$$WKHM(\mathrm{D},\mathrm{M}) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \sum_{f\in\{c,\rho\}} w_f \left\| d_{if} - m_{jf} \right\|^{-p}} \tag{2}$$

Even exponent p values are suggested, in our implementation we chose p=4.

B. Normalized Spectral Clustering

As an alternative to the *k*-means clustering algorithm family, we select to implement a version of normalized spectral clustering [15], which is similar to the Normalized Cut algorithm implemented by Ji and Geroliminis [11]. The advantage of this approach is the fact that feature weights are unnecessary, because spectral clustering treats clusters as connected subgraphs, while the *k*-means algorithm family treats them as distinct convex sets [16]. Additionally, this approach ignores sparse interconnections among subgraphs, providing a better defined boundary.

We chose to represent the selected urban traffic network in two ways, first as an undirected graph $G = (V, E)$ where each node $i \in V$ represents a road link and each edge $e \in E$ represents an intersection. Each node is associated with a density value $\rho_i = N_i / L_i, \forall i \in V$, where N_i stands for number of vehicles within link i and L_i stands for length of link i. It has to be denoted that two way road link connections are considered separate undirected edges. This first representation is appropriate for the implementation of the *k*-means and *k*-harmonic means clustering algorithms. The second representation as a weighted undirected graph is better suited to the implementation of the normalized spectral clustering algorithm. The normalized spectral clustering algorithm can be described by the following steps:

1) The weights $\beta_e \in B \in \mathrm{R}^{n \times n}$ on the edges belong to the weighted adjacency matrix B and are described by the similarity function between link density values which takes the form of a Gaussian probability distribution function $\exp(-(\rho_i - \rho_j)^2 / 2\sigma^2),\ \forall i, j \in V$ which punishes density value differences, decreasing in value faster the larger the differences.

2) After obtaining the weighted adjacency matrix B, we need to calculate the degree matrix, a diagonal matrix $R \in \mathrm{R}^{n \times n}$, whose diagonal elements can be calculated as follows: $r_i = \sum_{i=1}^{n} \beta_{ij}, \forall i \in V$.

3) The normalized Laplacian matrix $L_s \in \mathrm{R}^{n \times n}$ is then calculated by $L_s = I - R^{-1/2} B R^{-1/2}$.

4) Subsequently, we calculate the first k generalized eigenvectors $y_1, ..., y_k$ by solving the equation $L_s y = \lambda R y$.

5) Normalization of the rows of matrix $Y \in \mathrm{R}^{n \times k}$ to L^1-norm is followed by implementation of k-harmonic means clustering to matrix $Y_s \in \mathrm{R}^{n \times k}$.

6) Finally we get a set of clusters $C = \{C_1, ..., C_k\}$ which correspond to the initial data points cluster membership.

C. *Performance metrics*

Ji and Geroliminis [11], based on metrics developed by Rousseeuw [17], derived metrics to measure the performance of the Normalized Cut algorithm in comparison with a weighted k-means clustering implementation and showed that for a set of clusters $C = \{C_1, ..., C_k\}$ we can measure the intra-cluster density variability, as well as the inter-cluster density variability conditioned on the adjacency of the clusters being compared. Thus we have

$$CV(\mathrm{C}_v) = \frac{2\,\mathrm{var}(C_v)}{\mathrm{var}(C_v) + \mathrm{var}(C_u) + (\mu_{C_v} - \mu_{C_u})^2} \tag{3}$$

where C_v is the cluster whose performance we want to measure and C_u is the adjacent cluster with the least inter-cluster variability as regards C_v. Small values for the variances denoted by var(C) and large difference between the means, denoted by μ_c, gives small values for *CV* which means the cluster partitioning is successful. It has to be noted that we take the average *CV* over all clusters to account for the difference in link density among adjacent clusters. In order to measure spatial cohesion for each cluster, we decided to use the silhouette coefficient developed by Kaufman and Rousseeuw [18], which, for each data point measures two quantities, cohesion $\alpha(x)$, which measures average distance between data points within the cluster and separation $b(x)$, which measures the minimum average distance of data points to other clusters. Then we define silhouette $s(x)$ as

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \tag{4}$$

where $s(x) \in [-1,1]$ with values close to -1 considered as spatially poorly connected and values close to 1 considered as spatially well connected. The silhouette coefficient is the average for all data points and can be calculated using $SC = \sum_{i=1}^{N} s(x_i) / N$. In addition to calculating the isolated performance for spatial cohesion, this metric can be adapted to measure the total performance of the partitioning, by simply selecting both spatial and density features for each data point x_i. We will use *TP* to represent this metric for the remainder of this paper. It should be noted that for extreme values on either end of the spectrum, *TP* should have a limited impact during the clustering algorithm comparison, since its value is skewed by the extremely low or high values.

D. *Network Set up*

As mentioned before, using prior information about hotspot areas throughout Singapore for specific time periods, we selected an urban region with an area of about 2.2 km^2, located to the south of the island of Singapore, encompassing the Marina Bay area, as well as parts of the Central Business District, Chinatown, Dhoby Ghaut, and Clarke Quay. Our network consisted of about 90 intersections and 180 links. It should be noted that bidirectional road segments are considered as separate links. Link lengths range from 100m to 1000m and the speed limit is set to 50km/h. We used loop detector data obtained from the Land Transport Authority in Singapore for the month of August 2010, and based on Aslam et al. [12] results regarding peak traffic volumes in that particular area, decided that we should implement our tests for a time period of 3 hours, starting from 8am and ending at 11am, based on historical data for the first Monday of August 2010. Instead of Cartesian coordinates, we use Geographic coordinates (decimal degrees) in our plots, with longitude on the x-axis and latitude on the y-axis. The network can be shown in Fig. 2.

Figure 2. Illustration of the urban traffic network representing the Marina Bay Area of Singapore, which includes parts of Chinatown, Clarke Quay, Dhoby Ghaut and the Central Business District

III. EXPERIMENTAL RESULTS

Our main objective was to produce clusters with low intra-cluster density variability and high inter-cluster density variability. This objective, however, must satisfy the constraint of spatial cohesion. Having a cluster with very low link density variability, which includes links with absolutely no connectivity or spatial cohesion is useless for what we are trying to achieve. Another goal for us was to find the minimum number of clusters required for a sufficiently homogeneous partitioning. We implemented several test runs using all three algorithms. More specifically, we tested for different numbers of clusters, for different feature weights and for different times within the selected period of 8am to 11 am. It was determined that the optimal number of clusters is 2, which is understandable, since the encompassing area is not large enough to warrant further partitioning. After several runs, we were able to determine feature weights suitable for keeping a balance between our primary objective and the satisfaction of the spatial cohesion constraint. Finally, for 2 clusters and a weight ratio w_c/w_ρ of 100/1, we compare all algorithms throughout the 8am to 11am time period.

A. Performance comparison for number of clusters

The number of clusters for which the partitioning performance was investigated ranges from 2-4. There were runs for k = 5, but the resulting 5th cluster was empty, which informs us of the upper bound for our investigation. Obviously, if the performance indices for 2 clusters were also too high as regards link density variability, the only conclusion is that our selected network was homogeneously congested in its entirety. Fortunately, this was not the case, which means the region and time period selection was not made poorly. Due to the inherent tendency of k-means to reach local minima, dependent on the initial cluster center selection, we get a range of values for each cluster. Hence, Table I contains the average performance metric values derived from the k-harmonic means implementation, which exhibits insensitivity to initial cluster center selection. It should be noted, for this particular investigation, that the average of the range of values mentioned above and the performance metric values taken from the k-harmonic means algorithm implementation differ very slightly.

TABLE I. CLUSTER NUMBER COMPARISON

Performance Metrics	***k-number of clusters***		
	k=2	***k=3***	***k=4***
Average CV	***0.8705***	*1.0487*	*0.8183*
Average SC	***0.4713***	*0.6475*	*0.3966*
Average TP	***0.0957***	*-0.3217*	*-0.721*

It is evident that 2 clusters partitioning performs better in total. We should note that we set upper thresholds for *CV* and lower thresholds for *SC* and *TP* in order to facilitate the comparison. If the average *CV* value exceeds 1.0000, the partitioning is considered unsuccessful due to high inter-cluster link density similarity. If the average *SC* or *TP* value is below 0.0000, the partitioning is considered unsuccessful with respect to spatial cohesion, or a combination of low spatial cohesion and high inter-cluster similarity.

B. Performance comparison for different feature weights

For the reason mentioned in the previous section, we elect to utilize the k-harmonic means algorithm to perform the comparison for different feature weights. Table II contains the average performance metric values for different feature weight ratios w_c/w_ρ , ranging from 10/1 to 150/1.

TABLE II. FEATURE WEIGHT COMPARISON

Performance Metrics	w_c/w_ρ				
	10	***25***	***50***	***100***	***150***
Average CV	*0.0626*	*0.0814*	*0.1530*	***0.8705***	*0.9720*
Average SC	*-0.0749*	*-0.0578*	*-0.0566*	***0.4713***	*0.4782*
Average TP	*0.8740*	*0.8728*	*0.8647*	***0.0957***	*0.0285*

Admittedly, the weighting for spatial cohesion is quite aggressive, starting from merely 10/1 and eventually reaching 10 times the initial ratio, but based on a combination of performance metric values and graphical evidence, it was deemed necessary so as to enforce the critical constraint of spatial cohesion. From the performance metric values we can easily observe that, had we neglected to include the isolated spatial cohesion metric, our weight ratio selection would have been off by a magnitude of 10.

C. Performance comparison for k-means, k-harmonic means and normalized spectral clustering algorithms

We investigated the performance of each algorithm for several time slots within the selected time period of 8am to 11am. It is interesting to observe that the partitioning produced from each algorithm implementation displays the progression of traffic congestion throughout the time period. In Table III we compare all three algorithms, but, it should be pointed out that the k-means produces several sets of results for each time slot, dependent on the initial cluster center selection. Therefore, we have decided to present values to showcase that good performance can be achieved, but the instability of the results should lead us to disregard them from the algorithm comparison in the general case. For the majority of the time slots tested, k-harmonic means clustering outperformed normalized spectral clustering. Fig.3 graphically presents the clusters produced for *k*-means, *k*-harmonic means and normalized spectral clustering respectively, for time slots 8am, 9am, 10am and 11am. The progression of traffic congestion throughout the selected time period can easily be observed. Observation of the clusters shows, that an "ebb and flow" effect is taking place for that specific time period, which is consistent with peak traffic volume observations of Aslam et al. [11].

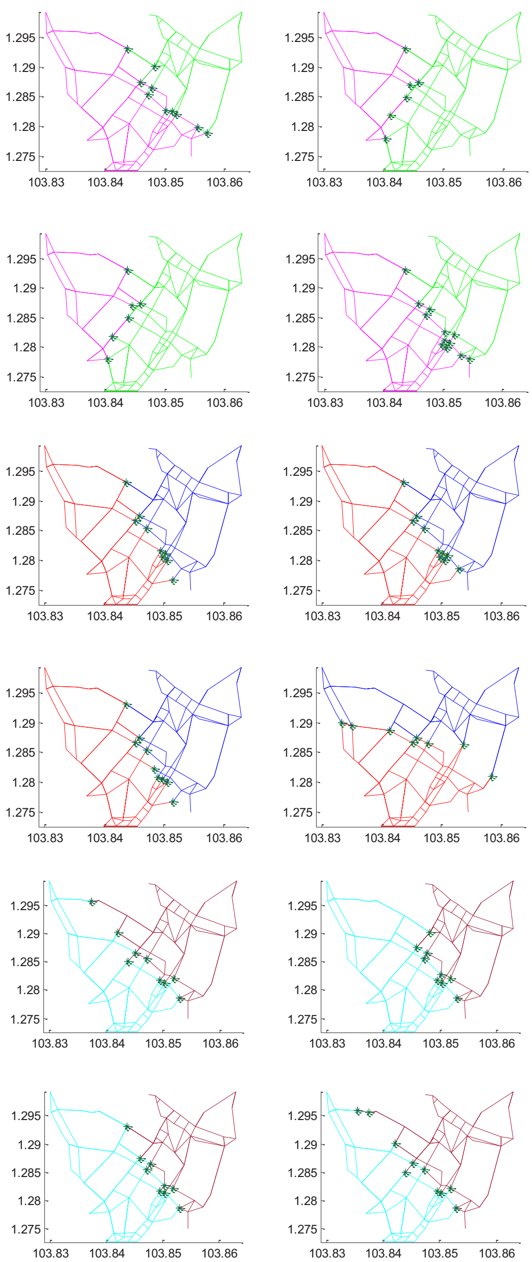

Figure 3. Network partitioning for time slots 8am (top left quadrant), 9am (top right quadrant), 10am (bottom left quadrant) and 11am (bottom right quadrant) using *k*-means (magenta-green), *k*-harmonic means (red-blue) and normalized spectral clustering (cyan-maroon) respectively.

TABLE III. CLUSTERING ALGORITHM COMPARISON

Performance Metrics	*Type of clustering algorithm*		
	WKM	***WKHM***	***NSC***
Average CV	*0.9237*	***0.9125***	*0.9451*
Average SC	*0.4753*	***0.4743***	*0.4703*
Average TP	*0.1555*	***0.0706***	*0.0305*

IV. CONCLUSIONS

From the experimental results, several conclusions can be reached. First, that the minimum number of clusters is dependent on the selected network topology and size. Second, that center-based clustering algorithms can be very effective in partitioning urban traffic networks into homogeneous regions, but stability remains an issue. Third, that spectral clustering approaches are suited to urban traffic network partitioning, omitting the tedious step of calculating optimal feature weights. Spatial cohesion is satisfied by the construction of the weighted adjacency matrix *B*. However, there is some computational effort involved in the adjacency matrix construction, which may lead to spectral clustering approaches being slower than center-based clustering approaches such as *k*-means and *k*-harmonic means. Our results also show that the weighted version of *k*-harmonic means outperforms normalized spectral clustering, as regards link density variability and spatial cohesion. One must consider however that in order to achieve the same level of spatial cohesion, very aggressive weighting of the coordinates feature had to be employed. Finally, after examining the different clusters produced from the clustering algorithms, we can deduce how congestion is formed and propagated over time. This pertains to historical data, which might not be reflecting the current traffic situation for that particular area, however this approach can be extended to apply to real-time data, or a combination of historical and real-time data. These approaches can be used as a middle layer of traffic surveillance, enabling traffic engineers to observe in real-time the traffic state, without overloading them with information. They can also serve as a first step in the derivation of the Urban-Scale Macroscopic Fundamental Diagram for each separate cluster, thus enabling the application of different perimeter control strategies along the boundary of the clusters.

ACKNOWLEDGMENT

The authors would like to thank Asst. Prof. Justin Dauwels for his help with data preprocessing and the Land Transport Authority of Singapore for providing the loop detector data.

REFERENCES

[1] "White Paper - European transport policy for 2010: time to decide", European Commission, 2001.

[2] J.W. Godfrey, "The mechanism of a road network", Traffic Engineering and Control, vol. 11, no 7, pp. 323–327, 1969.

[3] R. Herman and I. Prigogine. "A two-fluid approach to town traffic.", Science, vol 204, no 4389, pp. 148-151, 1979.

[4] R. Herman and S.A. Ardekani. "Characterizing traffic conditions in urban areas", Transportation Science, vol 18, no 2, pp. 101-140, 1984.

[5] C. F. Daganzo, "Urban gridlock: macroscopic modeling and mitigation approaches", Transportation Research Part B, vol 41, no 1, pp. 49-62, 2007.

[6] N. Geroliminis and C.F. Daganzo, “Macroscopic modeling of traffic in cities”, Transportation Research Record 86th Annual Meeting, Washington, D.C., no 07-0413, 2007.

[7] N. Geroliminis and C.F. Daganzo, “Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings”, Transportation Research Part B, vol 42, no 9, pp. 759-770, 2009.

[8] N. Geroliminis and J. Sun, “Properties of a well defined macroscopic fundamental diagram for urban traffic”, Transportation Research Part B, vol 45, no 3, pp. 605-617, 2011.

[9] A. Mazloumian, N. Geroliminis and D. Helbing, “The spatial variability of densities as determinant of urban network capacity”, Royal Society of London Philosophical Transactions Series A 368, pp. 4627-4647, 2010.

[10] C.F. Daganzo, V.V. Gayah and E.J. Gonzales, “Macroscopic relations of urban traffic variables: bifurcations, multivaluedness and instability.”, Transportation Research Part B, vol 45, no 1, pp. 278-288, 2011.

[11] Y. Ji and N. Geroliminis, “On the spatial partitioning of urban transportation networks.”, Transportation Research Part B, vol 46, no 10, pp. 1639-1656, 2012

[12] J. Aslam, S. Lim, X. Pan and D. Rus, “City-scale traffic estimation from a Roving Sensor Network.”, Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems, pp. 141-154. ACM, 2012.

[13] G. Hamerly and C. Elkan. "Alternatives to the k-means algorithm that find better clusterings.", Proceedings of the eleventh international conference on Information and knowledge management, pp. 600-607. ACM, 2002.

[14] B.Zhang, M. Hsu and U. Dayal, “K-harmonic means – a data clustering algorithm.”, Technical Report HPL-1999-124, Hewlett-Packard Labs, 1999.

[15] A. Ng, M. Jordan and Y. Weiss, “On Spectral Clustering Analysis and an algorithm.", Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press 14 2001, pp. 849-856, 2001.

[16] U. Von Luxburg, “A tutorial on spectral clustering.”, Statistics and computing, vol 17, no. 4, pp. 395-416, 2007.

[17] P. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.”, Journal of computational and applied mathematics vol 20, pp. 53-65, 1987.

[18] L. Kaufman and P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley and Sons, London, 1990.