hEART 2020 – 9th Symposium of the European Association for Research in Transportation
September 1–3, 2020, École nationale des travaux publics de l'État (ENTPE), Lyon, France

Demand estimation and spatio-temporal clustering for urban road networks

Chunli Zhu[1,2*] Jianping Wu[2] Anastasios Kouvelas[1]

[1] Traffic Engineering Group, Institute for Transport Planning and Systems
Swiss Federal Institute of Technology (ETH) Zurich, Switzerland
[2] Department of Civil Engineering,Tsinghua University, Beijing, China.
[*]Corresponding author

Abstract

Urban road networks play a key role in mobility amongst the critical infrastructure of city, which is a strong time-variant system with uncertainty. In this paper, for the purpose of understanding the traffic congestion propagation patterns, demand estimation and spatio-temporal clustering was performed with a case study on the central area of Nanjing, China. Firstly, a four-hour time-dependent origin-destination traffic demand is calibrated by utilizing the Adaptive Fine-tuning (AFT) algorithm, and aiming at minimizing the error between microscopic simulated results by SUMO and real-world Radio Frequency Identification (RFID) data. Then, spatio-temporal clustering was performed to illustrate the dynamic feature on congestion propagation by implementing the spectral clustering approach. Results demonstrate that the calibrated dynamic origin-destination matrix can illustrate a good match with RFID data, and the proposed spectral clustering approach is fast and feasible for the partitioning of urban road network.
Keywords: Demand estimation; spatio-temporal; spectral clustering; urban road network.

Introduction

Urban roadway systems play a vital role as the backbone and non-separable part of a city which provides assess and mobility to people, goods, and services. Essentially, with the rapid urbanization globally, the unbalance of demand and supply (i.e., road infrastructure), which in consequence creates traffic congestion, is experienced by all residents in everyday commuting. Meanwhile, in recent years, the fast development of Intelligent Transportation System (ITS), which is within the framework of smart city, provides a series of ways to help alleviate the negative effects of traffic congestion. If we consider Nanjing city, for instance, which is the second largest city in the southeast part of China, recently it was made mandatory for all newly licensed vehicles to install RFID tags. The penetration rate during the investigation period of this paper is approximately up to 40%. Although we can obtain "big data" everyday, the limited installation of RFID readers still makes it difficult to obtain the overall status of the urban road network. Therefore, it is not easy to analyze congestion propagation pattern only with the available RFID data sets, and further methodological tools and developments are needed for that purpose.

Traffic demand is immeasurable but can be estimated by various methods, with researchers having started to use link flow to estimate origin-destination (OD) matrices since the 1970s. As one of the most important parameters of traffic simulation models, traffic demand (i.e., OD matrix) is not easy to be calibrated for large-scale networks. Existing methods in literature can be classified as follows: (1)

bi-level programming models: traffic assignment models such as stochastic user equilibrium (SUE) is generally chosen as the lower level problem, which can consider the impact of congestion on path selection as well as the effect of the delay function; albeit, it has the inherent shortcoming on model computational complexity (Maher et al., 2001); (2) **statistical models**: this type of models provide a good pathway to combine the prior knowledge on OD matrix with current observations of traffic flow, for instance, maximum likelihood models (Spiess, 1987) and Bayesian models (Li, 2005); however, they are sensitive to the prior definition of OD matrices; (3) **simultaneous perturbation stochastic approximation (SPSA)**: SPSA and its variations have been applied extensively to such problems, due to their ability of simultaneously calibrating demand and supply parameters; nevertheless, they present a performance deterioration issue in terms of convergence, especially for large-scale networks (Antoniou et al., 2015, Tympakianaki et al., 2015). In conclusion, traffic demand estimation is not a new problem, but is still challenging when tackling a large-scale traffic system that has information of sparse traffic flow observations.

The understanding of congestion propagation patterns is critical, especially regarding its spatio-temporal relevance and transitivity (Yang and Wang, 2019), which is one of the most fundamental issues for the development of dynamic control strategies. Generally speaking, there exist three main types of studies focusing on this problem, which are: (1) **simulation modeling**: this type of models can emulate the system evolution dynamics to different dimensions dependent on the model itself, such as cellular transmission model, SIR (Susceptible, Infected, Recovered) model, car-following model and microscopic traffic simulation. Meanwhile, the set of parameters in a simulation model is of need great importance; (2) **data-mining**: it usually focuses on finding the causality among congestion, such as causal congestion trees (CCTs), frequent congestion subtree discovery, and dynamic Bayesian networks (DBN) (Nguyen et al., 2016). Although this type of methods can explain the causal relationship well, they usually make some assumptions (e.g. the impact of congestion is spatially independent); this can be suitable for local congestion propagation analysis, however, the isolation of continuous roads cannot reveal complete spatial transitivity; (3) **spatio-temporal clustering**: clustering algorithms are commonly used in works related to community detection (Malliaros and Vazirgiannis, 2013), image segmentation (Naik and Shah, 2014), sensor networks (Katiyar et al., 2010), and many others. The difficulties of urban road networks clustering mainly rely on its dynamic and uncertainty features that call for satisfactory resolution in both spatio-temporal dimensions. An initial segmenting/merging/boundary adjustment mechanism to minimize the variance of link densities while maintaining the spatial compactness of clusters was proposed by (Ji and Geroliminis, 2012). Later on, Saeedmanesh and Geroliminis have utilized a symmetric non-negative matrix factorization (SNMF) to assign links to proper clusters with high intra-similarity and low inter-similarity (Saeedmanesh and Geroliminis, 2016).

Essentially, a similarity function can be defined between traffic observations and significance testing can be utilized to tackle this problem. Such an approach, uses hypothesis testing to determine statistically significant clusters and perform the spatio-temporal clustering. In summary, for cases with limited traffic count data, the combination of a simulation model and spatio-temporal clustering could be of important value for practical cases. In the current paper, the adaptive fine tuning (AFT) algorithm has been utilized for offline traffic demand estimation with the Traffic Control Interface (TraCI) SUMO (Krajzewicz et al., 2002). A case study of the urban network of Nanjing, in China, was investigated. Our results show good performance and quick convergence for a large-scale system. The spectral clustering approach was employed for the spatio-temporal clustering, and simulation results demonstrate its high efficiency for real-time clustering. Conclusions and future work are discussed at the last part of the paper.

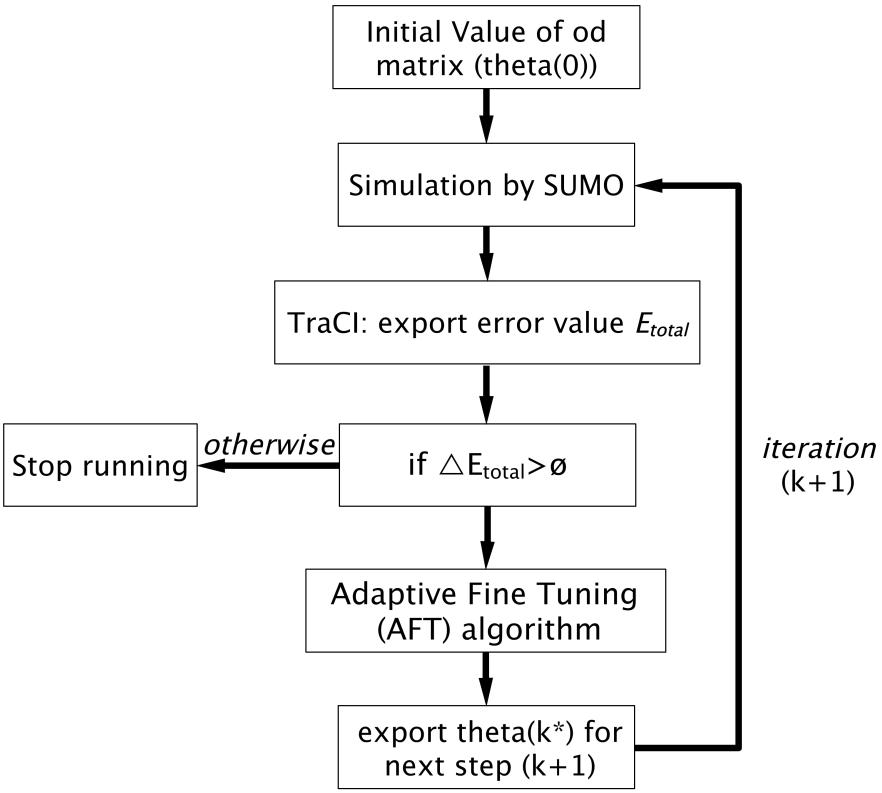


Figure 1: Flowchart of demand estimation framework.

Methodological framework

Figure 1 presents the flowchart of the demand estimation framework used in this paper. The TraCI API of the microscopic simulation model SUMO has been utilized to implemented the whole process. The input and output parameters θ represent elements of OD matrices, with $\theta(0)$, $\theta(*)$, and $\theta(k*)$ denoting the initial, best, and best values at k-th iteration, respectively. We denote with E_{total} the total error between the real and simulated data, and when the for difference between two consecutive iterations $\Delta E_{\text{total}} \leq \Phi$ is achieved, with Φ being a small positive number, the demand estimation process will be terminated. Here, the error E_{total} is defined as

$$E_{\text{total}} = \sqrt{\sum_{k}^{K} \sum_{i}^{N} (\hat{x}_i(k) - \bar{x}_i(k))} \tag{1}$$

where $\bar{x}_i(k)$ is the RFID data from reader i and $\hat{x}_i(k)$ is the SUMO simulated flows obtained by installed detectors, for the same time interval k; N denotes the number of RFID readers (i.e., detectors) and K the total discrete time-steps for all simulation horizon.

Briefly, the main concept of AFT algorithm is to use a universal approximator $\hat{J}(\theta, x)$ to obtain a modeling of the non-linear mapping $\hat{J}(\theta, x) \equiv J(\theta, x)$ between the free parameters and system output J; then an online adaptive/learning mechanism is employed to train the approximator with the iterative collected date set. At each iteration k, several random candidate perturbations are created for $\theta(k)$, while the best perturbation is selected based on the approximator $\hat{J}$; these correspond to the new tunable parameter values θ_{k+1} for the net time-step $k+1$ (entries of OD matrices in our case). It should be noted that AFT constitutes a generalization of SPSA algorithm, aiming at tuning system

parameters through iterations and learning of nonlinear system dynamics. The interested reader is referred to the literature (Kosmatopoulos et al., 2007; Kosmatopoulos and Kouvelas, 2009; Kouvelas et al., 2011(a); Kouvelas, 2011; Kouvelas et al., 2011(b)) for a detailed algorithmic description of AFT methodology.

Furthermore, for the spatio-temporal clustering, the spectral clustering approach is adopted in this paper to obtain a high inter-cluster and low intra-cluster similarity. This method is rather computationally efficient, especially if the affinity matrix is sparse. Spectral clustering uses information from the eigenvalues (spectrum) of special matrices built from the graph or the data set. For a graph $G(V, E)$, V and E denote the sets of vertices and links, respectively. In order to have a group of spatially connected links, the similarity function $w(i, j)$ between link i and link j is defined as follows

$$w(i,j) = \begin{cases} \exp(-(d_i - d_j)^2), & r(i,j) = 1 \\ 0, & r(i,j) > 1 \end{cases} \tag{2}$$

where we use the occupancy d of each link for expressing the similarity function. Distance $r(i, j)$ is calculated based on the adjacent matrix of graph $G(V, E)$, and i, j denote any couple of connected links; when they are adjacent $r(i, j) = 1$, and otherwise $r(i, j) > 1$ denotes their distance. The general steps to conduct the process of spectral clustering are the following:

(1) Construct the degree matrix D, and similarity matrix W.

(2) Compute the Laplacian graph $L = D - W$.

(3) The normalized Laplacian is given by $D^{-1/2}LD^{-1/2}$.

(4) Compute the first k eigenvectors $v_1, \ldots, v_k$ of L.

(5) Let $V \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $v_1, ..., v_k$ as column-wise.

(6) For $i = 1, ..., n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of V.

(7) Cluster the points $(y_i)_{i=1,...,n}$ in $\mathbb{R}^k$ with k-means algorithm into clusters $C_1, ..., C_k$.

More details on this approach can be found in Von Luxburg, 2007.

Case study and simulation results

The investigation area of this case study is within the highway boundaries colored in yellow in Figure 2, where 41 RFID readers data are available. This part is the most busy area in Nanjing city, which is the second largest city in southeast China. The investigation period is 6:00am to 10:00am on a working day; induction loop detectors were installed at the same locations with RFID readers in our microscopic simulation model in SUMO. The interval for exporting simulation data is set to 5 minutes, which is equal to the RFID data gathering frequency.

A 4-hour time-dependent demand estimation procedure was initially applied to the study network. For large networks with high dimensions, the number of links with available real-world data is usually much less than the number of OD pairs, which means that accurate solution of OD matrix cannot be obtained as the problem is under-determined. In Table 1, we report the total error values for the 4-hours dynamic demand, which show the performance performance of our estimation. Then, in Figure 3 we present a comparison of the error rates for individual detectors, which is defined as

$$e_i = \left| \frac{\hat{x}_i - \bar{x}_i}{\bar{x}_i} \right| \cdot 100\% \tag{3}$$

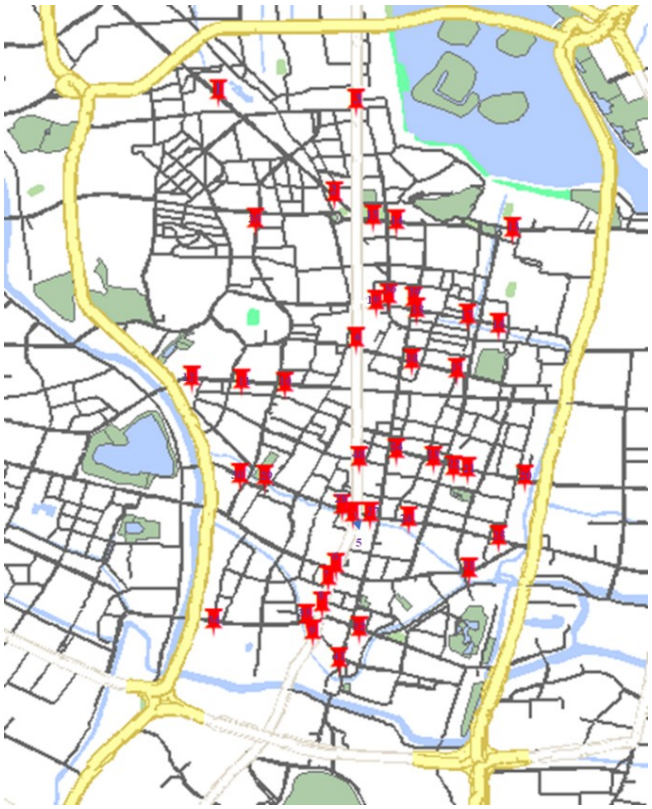


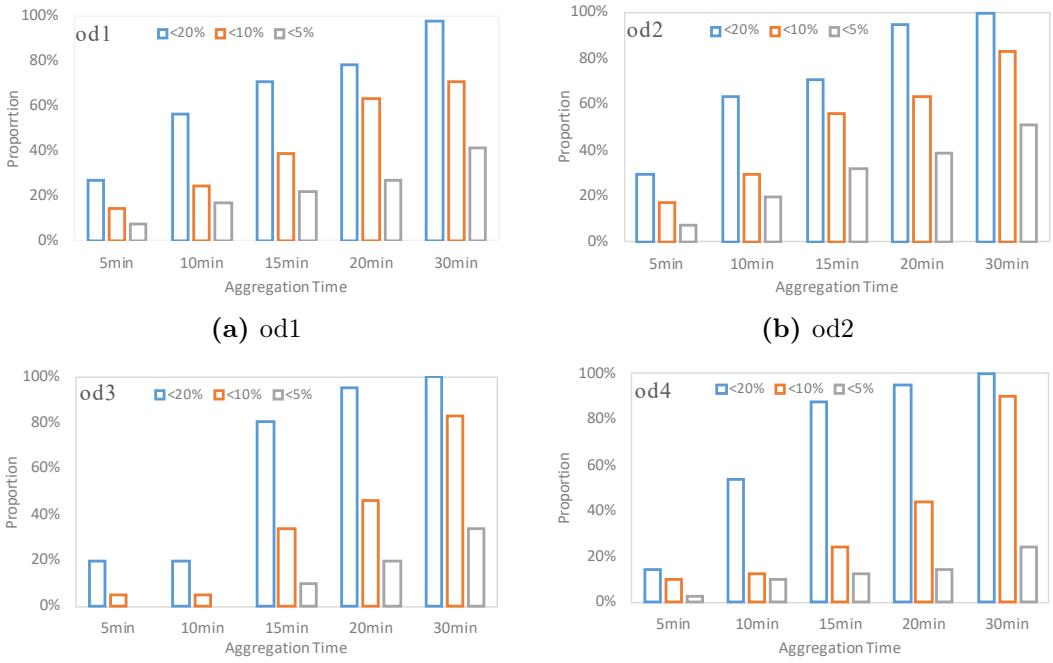
Figure 2: Investing area and positions of RFID readers.

Table 1: Error value of dynamic demand estimation by AFT.

OD	Time Period	Total Error
od1	6:00-7:00	990.22
od2	7:00-8:00	2586.29
od3	8:00-9:00	3009.48
od4	9:00-10:00	2808.97

What we observe from the 4-hours simulation results with the dynamic demand, is that increasing the aggregation time interval will lead to significant increase of the accuracy of results. As shown in Figure 3, for each case with aggregation time of 15 min, approximately 80% of links demonstrate error rate smaller than 20%. This is mainly due to the fact that averaging can eliminate the influence of volatility in simulation. Furthermore, the comparison between simulation and RFID data for some typical detectors are shown in Figure 4, in which 15 min was chosen as aggregation time. From Figure 4, it can be seen that simulation results of flow variables (veh/h) can present a satisfactory matching with RFID data, despite some few outliers. Note that these outliers could be readily eliminated using some sliding average method. The average error values of these three detectors during the 4 hours of simulation are presented in Table 2.

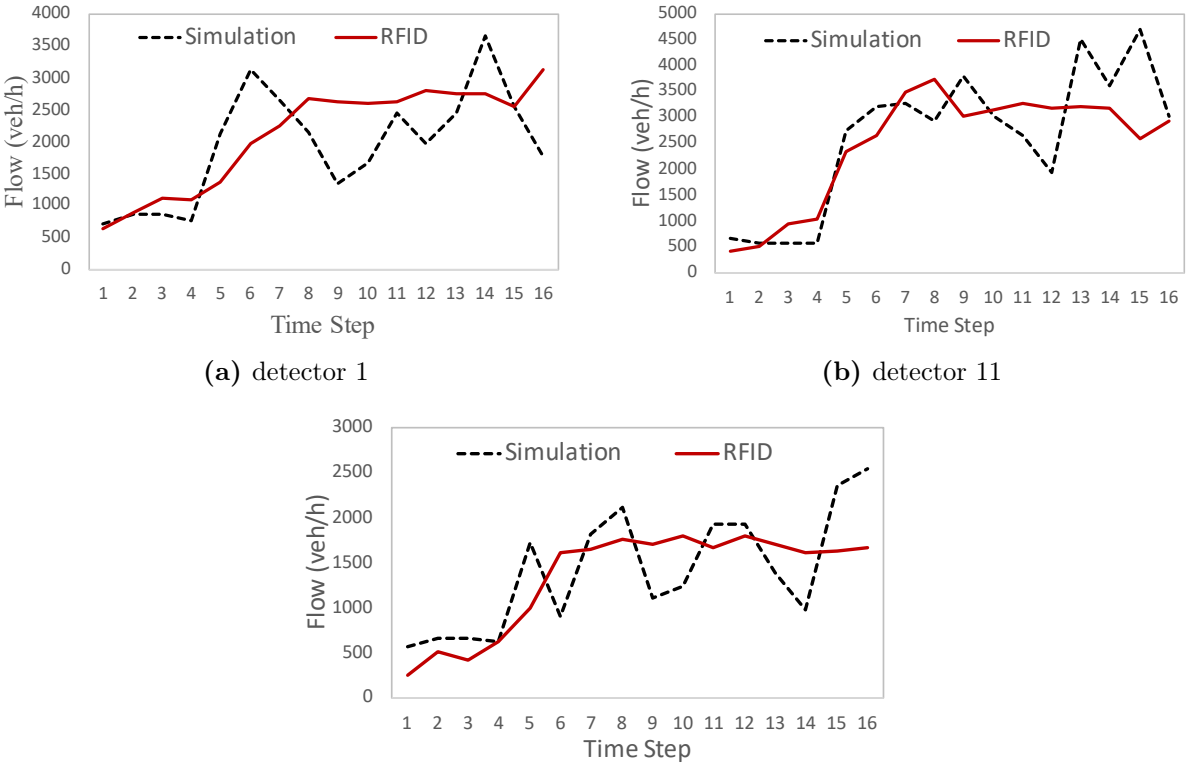
For the part of clustering, we built 93×93 adjacency matrix and similarity matrix; they are both sparse matrices. In Figure 5, some results of spatio-temporal clustering are presented, in which Figure 5(a)-(d), 5(e)-(h), 5(i)-(l), and 5(m)-(p), are averages between 6:15am-6:30am, 7:15am-7:30am, 8:15am-8:30am, and 9:15am-9:30am, respectively. Results with different number of clusters 3, 4, 5,



(a) od1 **(b)** od2

(c) od3 **(d)** od4

Figure 3: Proportion of detectors errors rate with different aggregation time.



(a) detector 1 **(b)** detector 11

(c) detector 16

Figure 4: Comparison between simulation and RFID data with aggregation time 15 min.

Table 2: Average error value during 4 hours time horizon.

	6:00-7:00	7:00-8:00	8:00-9:00	9:00-10:00
detector 1	-3.56%	9.16%	-10.04%	-1.83%
detector 11	-0.49%	0.78%	-3.07%	11.57%
detector 16	18.22%	5.12%	-2.80%	0.20%

and 6, respectively, are demonstrated. In order to compare the quality of the partitioning, we use the total variance of the clusters, which can be expressed as

$$\mathrm{NS}_k = \frac{\mathrm{NS}_k(A,A)}{\mathrm{NS}_k(A,B)} = \frac{2\mathrm{Var}(A)}{\mathrm{Var}(A)+\mathrm{Var}(B)+(u_A-u_B)^2} \tag{4}$$

where A and B are neighbouring clusters, $\mathrm{Var}(\cdot)$ is the variance of each cluster, and u denotes the observation index (e.g., occupancy has been used here). Then the average NS values for each time interval and different number of cluster are compared as shown in Table 3. When there is no congestion in the network (e.g. 6:15am-6:30am), results with 3 clusters present the minimum variance, while for some cases with congestion (e.g. 8:15am-8:30am) the 5 clusters have the best result. From these results, it can be reflected that traffic congestion improves the heterogeneity of the road network.

For a clearer understanding of the congestion propagation process, we also calculate the statistical results of the partitioning that present the best average NS (Table 4). Combined with Figure 5, a clear view of congestion propagation from the blue cluster in 5(f) to a larger area, the red cluster in 5(k), and then congestion shows a degrading trend as the red cluster in 5(o). During the morning peak time, namely 8:15am to 8:30am, the number of links in the red cluster is the largest.

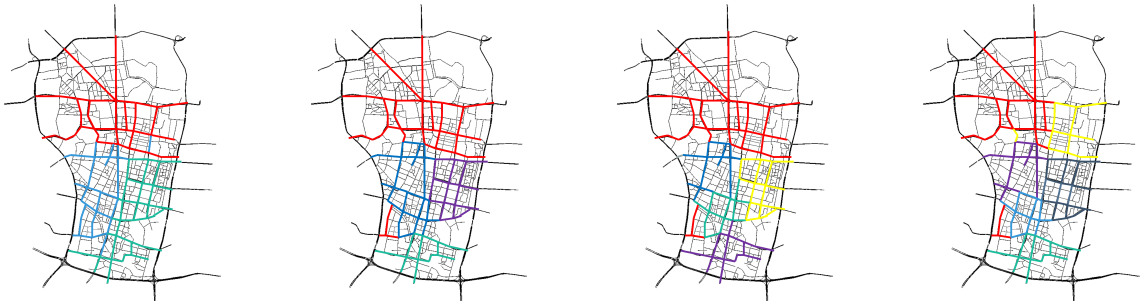
Conclusions and future work

In this paper, we have discussed the demand estimation and spatio-temporal clustering for large-scale urban road networks. For the former part, an iterative framework has been implemented with TraCI API in SUMO, and AFT algorithm has been utilized for determining the best OD dynamic matrices. Results of this framework show good matching with real-world RFID data collected from the network, and higher aggregation intervals illustrate better results. For the latter part, the spectral clustering approach has been chosen due to its good performance when dealing with sparse similarity matrices. The presented results have demonstrated that spectral clustering is a feasible and efficient pathway for investigating urban road networks spatio-temporal characteristics. Combined with statistical data for different number of clusters, one can study congestion patterns, i.e. generation, propagation, and degradation in a large city-wide context.

In the future, we will focus on spatio-temporal clustering when dealing with bi-directional flows. Furthermore, signal control strategies will be studied for the prevention of congestion propagation.

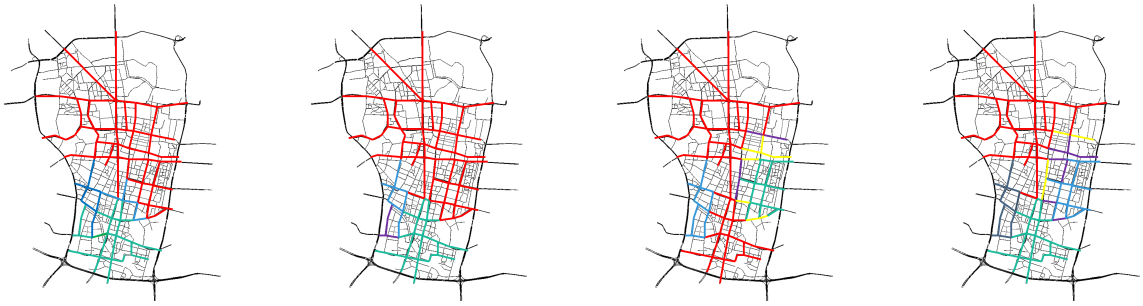
Table 3: Average NS value of different cluster.

Number of Clusters	2	3	4	5	6
6:15-6:30	0.999	**0.830**	0.971	0.997	0.924
7:15-7:30	0.990	0.829	**0.806**	0.812	0.870
8:15-8:30	0.990	0.957	0.945	**0.853**	0.941
9:15-9:30	0.975	0.966	0.994	**0.958**	1.137

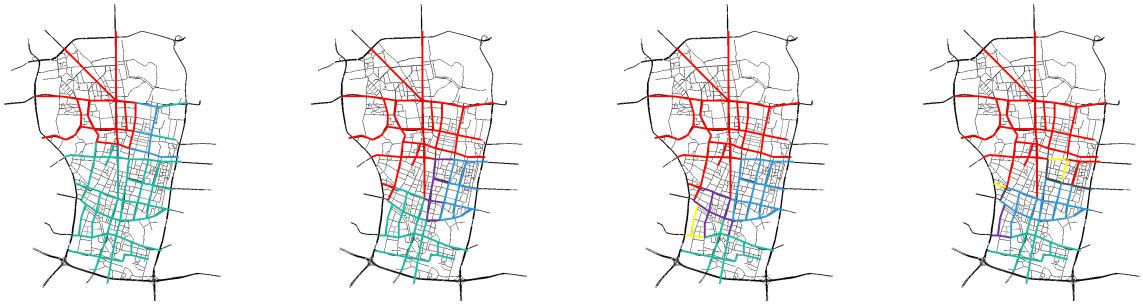


(a) (6:15-6:30)-3 cluster **(b)** (6:15-6:30)-4 cluster **(c)** (6:15-6:30)-5 cluster **(d)** (6:15-6:30)-6 cluster

(e) (7:15-7:30)-3 cluster **(f)** (7:15-7:30)-4 cluster **(g)** (7:15-7:30)-5 cluster **(h)** (7:15-7:30)-6 cluster



(i) (8:15-8:30)-3 cluster **(j)** (8:15-8:30)-4 cluster **(k)** (8:15-8:30)-5 cluster **(l)** (8:15-8:30)-6 cluster



(m) (9:15-9:30)-3 cluster **(n)** (9:15-9:30)-4 cluster **(o)** (9:15-9:30)-5 cluster **(p)** (9:15-9:30)-6 cluster

Figure 5: Results of spatio-temporal clustering

cluster	red	blue	green
Average occupancy	0.156	0.213	0.154
Variance	0.020	0.036	0.016
Number of links	32	26	35

(a) 3 clusters during 6:15am-6:30am (Figure 5(a))

cluster	red	blue	green	purple
Average occupancy	0.349	0.466	0.207	0.270
Variance	0.086	0.107	0.026	0.046
Number of links	22	23	32	16

(b) 4 clusters during 7:15am-7:30am (Figure 5(f))

cluster	red	blue	green	purple	yellow
Average occupancy	0.429	0.293	0.288	0.273	0.336
Variance	0.210	0.078	0.062	0.009	0.029
Number of links	58	10	13	4	8

(c) 5 clusters during 8:15am-8:30am (Figure 5(k))

cluster	red	blue	green	purple	yellow
Average occupancy	0.462	0.293	0.171	0.316	0.264
Variance	0.138	0.052	0.037	0.048	0.133
Number of links	45	22	14	9	3

(d) 5 clusters during 9:15am-9:30am (Figure 5(o))

Table 4: Statistical results of spatio-temporal clustering.

References

Antoniou, C., Azevedo, C. L., Lu, L., Pereira, F. and Ben-Akiva, M. (2015). W-spsa in practice: Approximation of weight matrices and calibration of traffic simulation models, *Transportation Research Part C: Emerging Technologies* **59**: 129–146.

Ji, Y. and Geroliminis, N. (2012). On the spatial partitioning of urban transportation networks, *Transportation Research Part B: Methodological* **46**(10): 1639–1656.

Katiyar, V., Chand, N. and Soni, S. (2010). Clustering algorithms for heterogeneous wireless sensor network: A survey, *International Journal of applied Engineering research* **1**(2): 273.

Kosmatopoulos, E. B. and Kouvelas, A. (2009). Large scale nonlinear control system fine-tuning through learning, *IEEE Transactions on Neural Networks* **20**(6): 1009–1023.

Kosmatopoulos, E. B., Papageorgiou, M., Vakouli, A. and Kouvelas, A. (2007). Adaptive fine-tuning of nonlinear control systems with application to the urban traffic control strategy tuc, *IEEE Transactions on Control Systems Technology* **15**(6): 991–1002.

Kouvelas, A. (2011). Adaptive fine-tuning for large-scale nonlinear traffic control systems (phd thesis), *Technical University of Crete, Chania, Greece* .

Kouvelas, A., Aboudolas, K., Kosmatopoulos, E. B. and Papageorgiou, M. (2011(b)). Adaptive performance optimization for large-scale traffic control systems, *IEEE Transactions on Intelligent Transportation Systems* **12**(4): 1434–1445.

Kouvelas, A., Papageorgiou, M., Kosmatopoulos, E. B. and Papamichail, I. (2011(a)). A learning technique for deploying self-tuning traffic control systems, *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 1646–1651.

Krajzewicz, D., Hertkorn, G., Rössel, C. and Wagner, P. (2002). Sumo (simulation of urban mobility)-an open-source traffic simulation, *Proceedings of the 4th middle East Symposium on Simulation and Modelling (MESM20002)*, pp. 183–187.

Li, B. (2005). Bayesian inference for origin-destination matrices of transport networks using the em algorithm, *Technometrics* **47**(4): 399–408.

Maher, M. J., Zhang, X. and Van Vliet, D. (2001). A bi-level programming approach for trip matrix estimation and traffic control problems with stochastic user equilibrium link flows, *Transportation Research Part B: Methodological* **35**(1): 23–40.

Malliaros, F. D. and Vazirgiannis, M. (2013). Clustering and community detection in directed networks: A survey, *Physics Reports* **533**(4): 95–142.

Naik, D. and Shah, P. (2014). A review on image segmentation clustering algorithms, *Int J Comput Sci Inform Technol* **5**(3): 3289–93.

Nguyen, H., Liu, W. and Chen, F. (2016). Discovering congestion propagation patterns in spatio-temporal traffic data, *IEEE Transactions on Big Data* **3**(2): 169–180.

Saeedmanesh, M. and Geroliminis, N. (2016). Clustering of heterogeneous networks with directional flows based on "snake" similarities, *Transportation Research Part B: Methodological* **91**: 250–269.

Spiess, H. (1987). A maximum likelihood model for estimating origin-destination matrices, *Transportation Research Part B: Methodological* **21**(5): 395–412.

Tympakianaki, A., Koutsopoulos, H. N. and Jenelius, E. (2015). c-SPSA: Cluster-wise simultaneous perturbation stochastic approximation algorithm and its application to dynamic origin–destination matrix estimation, *Transportation Research Part C: Emerging Technologies* **55**: 231–245.

Von Luxburg, U. (2007). A tutorial on spectral clustering, *Statistics and computing* **17**(4): 395–416.

Yang, L. and Wang, L. (2019). Mining traffic congestion propagation patterns based on spatio-temporal co-location patterns, *Evolutionary Intelligence* pp. 1–13.