Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

International Symposium of Transport Simulation (ISTS'18) and the International Workshop on Traffic Data Collection and its Standardization (IWTDCS'18)

Local traffic patterns extraction with network-wide consistency in large urban networks

Yaroslav Hernandez[a,*], Tamara Djukic[a], Jordi Casas[a]

[a]*Aimsun SL, Ronda de Universitat 22B, Barcelona 08007, Spain*

Abstract

With the increasing availability of traffic data in large urban areas, there is an opportunity to infer more sophisticated traffic patterns and trends which till now has been difficult to obtain and understand. The inferred traffic patterns can then serve as input into short-term traffic prediction models or to predict traffic demand in the network for deployment of traffic management policies. However, the network-wide consistency, the interaction between variables and the difficulty during forecasting in choosing a pattern at early AM peak hour require robust, but flexible enough pattern extraction method. A wide range of researchers tackle some of these points, but it is difficult to find the one that fulfills all of them.

Thus, this paper presents traffic flow pattern extraction methodology to identify daily traffic flow profiles consistent among all the detectors as a patterns based on two-phase methodology consisting of decision tree algorithm, Pathmox, and 2-step iterative clustering. The first phase uses qualitative variables (i.e., holidays, large events, weather, day-of-week) to capture meaningful and robust patterns in a tree-based configuration. The second phase consists of two steps that aim to reduce still high variability in some patterns non-attributable to the qualitative factors already exploited in the first phase.The advantages of this approach are (1) its capacity to break down variability in patterns due to both known and unknown factors; (2) it does not rely on specific network settings and (3) network-wide scale consistent patterns are identified. The traffic pattern extraction method is evaluated with real traffic flow data collected in period of one year on a motorway network M4 and M7 in Sydney, Australia. Results show that the proposed method extracts more identifiable patterns and more efficiently captures trends in the data with almost non-overlapping conditional variability bands compared to spatial-clustering approach based on *k*-means.

Keywords: Traffic flow pattern; Pathmox algorithm; network-wide consistency; large-scale.

* Corresponding author. Tel.: +34-933-171-693.
E-mail address: yaroslav.hernandez@aimsun.com

1. Introduction

With the increasing availability of traffic data in large urban areas, there is an opportunity to infer more sophisticated traffic patterns and trends which till now has been difficult to obtain and understand. The inferred traffic patterns can then serve as input into short-term traffic prediction models or to predict traffic demand in the network for deployment of traffic management policies.

Inference of patterns is not easy task due to the inherent variability in traffic behaviour and consistency requirements among all detectors, as the goal is to extract a minimum number of patterns with lowest conditional variability bands (Gasser et al. (1986)). Noticeable research efforts in the relevant literature focus on traffic patterns extraction (Soriguera (2012); Han and Moutarde (2016); Salamanis et al. (2017)), including a range of matrix factorisation techniques and clustering algorithms. For example, K-means as the most common used clustering algorithm due to its simplicity and efficiency, has been applied to uncover the hidden structures in huge traffic data sets for identifying daily congestion or traffic volume patterns (Yang et al. (2017)). Further, Yang et al. (2017) performed spectral clustering approach on speed data, obtaining a set of clusters where the data are a mix of sections and calendar dates. Additionally, they assume an isotropic distribution with $\sigma = 1$, which breaks temporal properties of daily traffic profiles. Also, different application motivations for modeling urban traffic led to diverse traffic pattern definitions in existing literature that mainly deals with finding traffic patterns at local, detector level without ensuring spatial consistence of local patterns over all the detectors in the network. For example, Attanasi et al. (2017) proposed a hybrid approach in the sense of integrating historical data patterns extracted by Affinity Propagation clustering method and real-time data as input in statistical model for short-term flow prediction. However, authors consider each detector as independent object and perform local clustering for each of them independently, which is not suitable for the purposes of network-wide traffic and demand prediction as the network consistency is an essential requirement. Also, Soriguera (2012) proposed a three-step clustering, 1) by day-of-week, 2) seasonal variations and 3) special days. The extracted patterns are section-dependent, that is, different sections may be affected by different sets of patterns and even same patterns may represent different sets of days. Hence, the consistency requirement is not fulfilled. Salamanis et al. (2017) performed segmentation of data and their clustering to fit more specific models for each combination leading to final NRMSE improvement in both normal and abnormal traffic conditions. It resembles the mixed models framework for longitudinal unit data. However, the extracted patterns correspond to one single location and different segments of daily flow profile. Therefore, it is not clear how to extend this methodology to a whole network preserving consistency among locations. Han and Moutarde (2013) proposed a non-negative matrix factorization to derive lower-dimensional representation of the network-level traffic states to further perform k-means clustering. They introduced the structural regularization term in the objective function in order to preserve geometrical structures of original data in the new low-dimensional representation. This results into two dimensional representation of data set, where columns represent a traffic state observation at a given simulation scenario and time-step, and rows represent these observations at corresponding detectors. Therefore, different simulation scenarios (days) or different time-steps may fall into the same clusters. To overcome this 2-dimensional limitation, Han and Moutarde (2016) applied tensor representation on a data set to extract patterns consistent among all detectors in the network. However, their method relies only on non-contextual information that may lead to poor quality of traffic prediction at early AM peak hour, because the choice would rely on less discriminative night hours.

However, the network-wide consistency, the interaction between variables and the difficulty during forecasting in choosing a pattern at early AM peak hour require robust, but flexible enough pattern extraction method. Interpretability power appears to be highly valuable when prediction fails, as then it is easier to detect whether the model is not appropriate anymore or prediction fails due to non-recurring conditions. Although, user pre-defined clusters are highly interpretable, they lack rigorous methodological approach and even manually testing different configurations may be very inefficient. A wide range of researchers tackle some of these points, but it is difficult to find the one that fulfills all of them. Thus, this paper presents developed traffic flow pattern extraction methodology to identify daily traffic flow profiles consistent among all the detectors as a patterns based on decision tree algorithm, Pathmox (Lamberti et al. (2016)), and 2-step iterative clustering. The advantages of this approach are (1) its capacity to break down variability in patterns due to both known and unknown factors; (2) it does not rely on specific network settings and (3) network-wide scale consistent patterns are identified.

The paper is organized as follows. In the first part of the paper, the network-wide consistent traffic pattern extraction methodology is presented that constitutes of two main phases, contextual and non-contextual clustering. Next, the main properties of each phase that ensure robust, non-overlapping and network-wide consistent patterns are explained in more details. In the last part of this paper, we demonstrate the performance of the proposed traffic pattern extraction method on a motorway network, M4 and M7, Sydney, Australia. The paper closes with a discussion on further application perspectives of the traffic pattern extraction method and further research directions.

2. Methodology

The proposed network-wide consistent traffic pattern extraction methodology constitutes two main phases, contextual and non-contextual clustering as presented in Fig. (1). The first phase uses qualitative variables (i.e., holidays, large events, weather, day-of-week, season) to capture meaningful and robust patterns in a tree-based configuration. The second phase consists of two steps that aim to reduce still high variability in some patterns non-attributable to the qualitative factors already exploited in the first phase (denoted in purple in Fig. 1). Hence, we refer to method in phase 2 as *hybrid* due to joint modelling of qualitative knowledge with unknown external factors that highly influence variability in the traffic flow data.

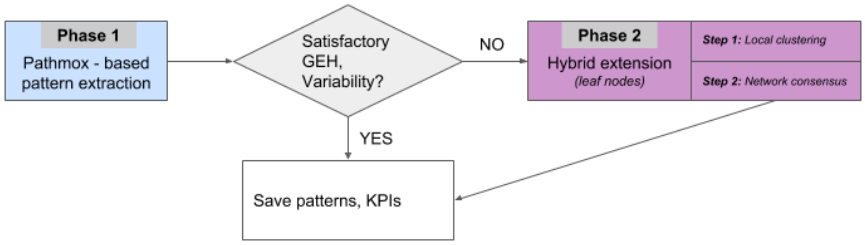


Fig. 1: Methodology for network-wide consistent traffic flow pattern extraction.

2.1. Phase 1: The Pathmox-based algorithm for contextual clustering

In the contextual clustering phase, pattern extraction based on heterogeneity modelling statistical technique Pathmox, (Lamberti et al. (2016)) is applied to extract patterns that capture the interpretable variability in the data due to seasonal, weather, or day-of-week changes. Pathmox algorithm is a decision tree algorithm that avoids combinatorial problems due to its greedy nature (Breiman et al. (1984)). Its main idea is to perform splits at each node of the tree by performing a statistical hypothesis tests to determine which binary qualitative variable (e.g, day of the week) partitions the daily traffic flow observations in such a way that their corresponding models' difference is most significant. We have extended Pathmox algorithm by introducing non-parametric models to extract the patterns. More specifically, local polynomial regression model with the corresponding bandwidth has been chosen to capture the non-linearity in data. Additionally, the entropy criteria is used to balance the tree structure to ensure robust and non-overlapping patterns. In this way, traffic flow patterns that constitute daily traffic profiles are extracted from data and labelled as for example a dayOfWeek-season-weather.

Thus, the hypothesis test is performed at each node over the binary partition of the data by each qualitative variable producing two sub-populations. Therefore, the non-parametric models of the form $y_{ij} = m_i(x_{ij}) + \varepsilon_{ij}$ with $i = \{1, 2\}$, $j = 1 \dots n_i$ are compared by means of the hypothesis test (Bowman and Azzalini (1997))

$$\begin{cases} H_0 : m_i = m, \quad i = \{1, 2\} \\ H_1 : \text{not all regression curves are equal} \end{cases} \tag{1}$$

The test statistic is formulated accordingly

$$s = \frac{\sum_{i\in\{1,2\}} \sum_{j=1}^{n_i} \left(\hat{m}_i(x_{ij}) - \hat{m}(x_{ij})\right)^2}{\hat{\sigma}^2} \tag{2}$$

where, $\hat{m}$ is the joint model under the null hypothesis and $\hat{m}_i$ is the sub-population specific model, and $\hat{\sigma}^2 = \frac{\sum_{i\in\{1,2\}} \eta_i \hat{\sigma}_i^2}{\sum_{i\in\{1,2\}} \eta_i}$, is coefficient with η_i an effective number of d.o.f., is the pooled estimate of $Var(\varepsilon_{ij})$.

Now, one way of tabulating the distribution of s, the ANOVA-like statistic, is by means of permutation test. The probability is calculated as $\frac{\#\{s_b > s_{\text{obs}}\}}{B}$ with $b = 1 \ldots B$, where B samples are generated with random permutation of sub-population labels.

2.2. Phase 2: Non-contextual clustering

Given a number n of qualitative patterns determined in phase 1, the non-contextual phase is responsible for a hybrid extension on tree nodes to reduce further variability in patterns linked to unknown external factors (see purple nodes in Fig. 3a). This phase consists of two steps: (i) local pattern clustering, and (ii) network consensus. In the first step, algorithm selects m nodes, where $m < n$, with the highest variability and performs clustering at each detector individually to further reduce variability. At this point, the main drawback of clustering is the non-consistency between combination of observations (i.e., dates) within pattern over all detectors. Hence, the second phase aims to iteratively rebalance daily traffic profiles at each detector such that they form patterns consistent for all the detectors at a time. This step is the most important part of the proposed methodology.

Let us assume there will be positive and negative cluster partitions G_+ and G_- of the total set of dates D in a given Pathmox' leaf node, having $G_+ \cap G_- = \varnothing$ and $G_+ \cup G_- \subseteq D$. Moreover, *within* and *between* inertia, denoted as $w_{d,s,g}$ and $b_{d,s,g}$ respectively, are defined as optimality criteria for a traffic flow observation $\boldsymbol{x}_{d,s}$ at a particular date d, detector s and set of dates g, as follows

$$w_{d,s,g} = \sqrt{\sum_{i}^{T} (\boldsymbol{x}_{d,s,i} - \boldsymbol{c}_{s,i}^{(g)})^2} \qquad \text{and} \qquad b_{d,s,g} = \sum_{i}^{T} |\boldsymbol{x}_{d,s,i} - \boldsymbol{c}_{s,i}^{(g)}| \tag{3}$$

where $\boldsymbol{c}_s^{(g)}$ is the average traffic flow over g dates at detector s and T is a set of time-steps in a daily flow profile. The within inertia is minimized in least squares sense (L_2-norm) and the between inertia is maximized as an absolute deviation (L_1-norm), as we pursue homogeneously separated clusters. We start by assigning seed daily flow pattern to positive partition set

$$G_+ \cup \arg\max_{d\in D} \frac{\sum_{s\in S} b_{d,s,g_{d,s}^{\mathsf{c}}}}{\sum_{s\in S} w_{d,s,g_{d,s}}}$$

where $g_{d,s}$ is a set of dates, which represent sensor optimal local cluster to which date d has been initially assigned and $g_{d,s}^{\mathsf{c}}$ its complement. At this point, G_+ has one element that maximizes the between inertia and minimizes the within inertia among all detectors with respect to their local partitions. The negative set $G_- = \varnothing$. Now the idea is to assign dates most similar to G_+ and dates to G_- that maximize separation between G_+ and G_- detectors' average traffic flow.

This is solved iteratively using

$$\arg\min_{d\in D} \sum_{s\in S} w_{d,s,G_+} \qquad \text{and} \qquad \arg\max_{d\in D} \sum_{s\in S} \sum_{i}^{T} |\boldsymbol{c}_{s,i}^{(G_+)} - \boldsymbol{c}_{s,i}^{(G_-\cup\{d\})}| \tag{4}$$

The number of at most $\lfloor |D|/2 \rfloor$ dates is assigned to each partition $G_{+/-}$ by means of euclidean distance from G_+ and G_- to detector's individual clusters. Finally, the obtained sets G_+ and G_- are the optimal combination in a greedy sense that fulfill the detector's traffic flow patterns consistency at the network-wide level.

It is worth noting that minimization and maximization formulations in (4) are given as a sum over all sensors, being a more appropriate summarization statistic than an average or a median, specially in high-flow scenarios, where the variability is usually greater. Consider the situation in which only few detectors present a desired very high optimization function values, distorting the mean and providing wrong partitions which only benefit these few detectors. On the other hand, the median function is neither suitable, as for instance, more than half of sensors may present very low-volume (i.e., associated to low variability) and play almost no role in optimization process, leading to impractical selection of the quantile.

3. Experimental setup

3.1. Network and traffic data

The proposed traffic pattern extraction methodology is evaluated for the M4 and M7 motorways in Sydney, Australia. Real traffic flow data set over one year period, corresponding to periods Jul, 2016 – Dec, 2016 and Jan, 2017 – Jun, 2017, has been chosen in order to account impact of seasonality and all public holidays. The network consists of 394 detectors spread uniformly at each 500 metres, as presented in Fig. 2. The raw traffic flow data was exhaustively pre-processed, removing inconsistent values, stuck observations, extreme outliers and out-of-bounds values. Additionally, missing values have been imputed and data was aggregated by 15 minute interval. A 24-hour flow profiles were considered despite redundancy introduced in extracted patterns at night time due to their low variability.

3.2. Assessment scenarios

In determining performance of the proposed method, two benchmark scenarios are defined to evaluate the extracted traffic flow patterns and their robustness:

- The first scenario evaluates the gain in qualitative expert knowledge setting, where proposed traffic pattern extraction method has been compared with fixed clustering approach. The fixed clustering approach requires user pre-defined set of daily flow patterns based on qualitative expert knowledge. Several configurations have been tested manually and the best one, consisting of 8 patterns (i.e., holiday and day-of-week), have been selected for benchmark scenario.
- The second scenario benchmarks the proposed non-contextual clustering with spatial clustering method. Network-wide consistent traffic flow patterns for the node h|47 (good weather - Spring/Fall - non-holiday - working days) are identified by proposed method and compared with spatial clustering method. The spatial clustering comes from the design matrix setting, in which the rows are calendar dates corresponding to the pattern h|47 and the columns are the daily flow profiles for all detectors, having that cell (i, j) has the flow of time-step $j \mod (24\cdot 4)$ for date i and detector $\lfloor j/(24\cdot 4) \rfloor$. Finally, there will be as many columns as time-steps in a day per number of detectors $(24 \cdot 4 \cdot 394)$. NIPALS (Wold et al. (1987)) algorithm is applied to extract the first $\min(rank(X), \#cols(X) \cdot 0.8)$ principal components, reducing the dimensionality and automatically treating the missing values, which in turn are the input to k-means clustering algorithm, with $k = 2$. Note that in any method presented or compared to in this work, the patterns are always sets of dates that have as high as possible *intra-similarity* and as low as possible *inter-similarity*.

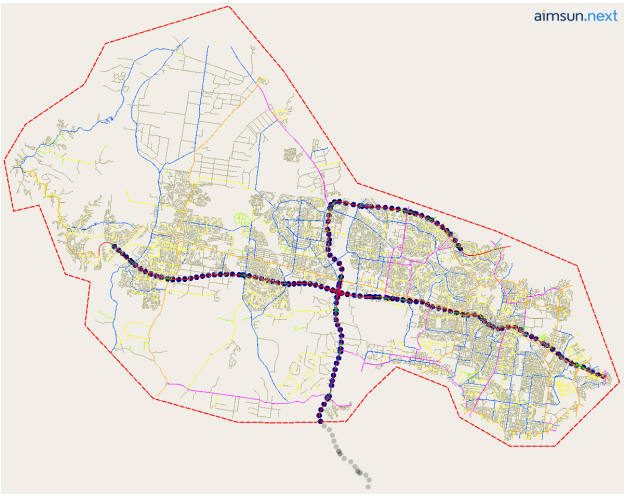
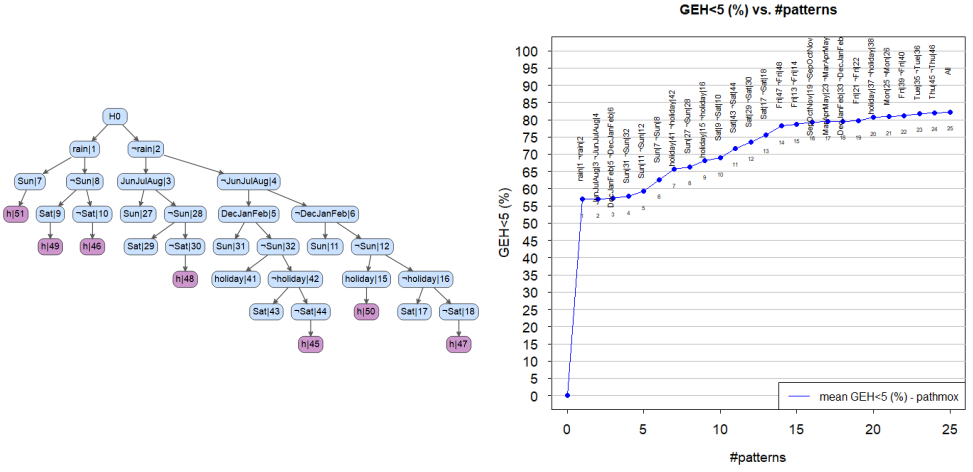


Fig. 2: M4-M7 Motorways (Sydney, Australia) where detectors are presented as a dark blue dots.

4. Results

Fig. 3a depicts the final tree-shape model derived using Pathmox-based, 2-phase method. There are 14 qualitative patterns (nodes) extracted, where 7 of them have a hybrid extension derived in phase 2 through iterative clustering. In general, the choice of the number of patterns that should be retained for phase 2 is often made by the visual examination of a number of different criteria. The simplest criterion relates to plotting of the GEH curve (Feldman (2012)) sorted by value and evaluation of the plot for an elbow. Fig. 3b presents the scree plot of GEH values obtained for increasing number of qualitative patterns. A sharp elbow appears at about the 14th patterns, which is consistent with condition to select a reasonably small number of qualitative patterns and improve the solution by applying the 2nd phase of the proposed approach.

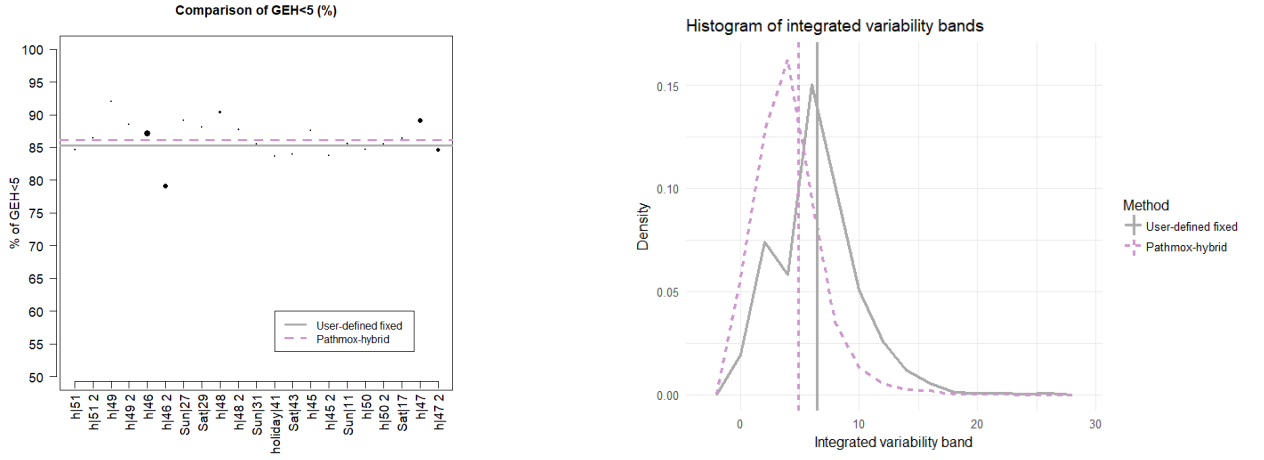


(a) Tree-shape model derived using Pathmox-based, 2-phase method

(b) GEH measure for increasing number of qualitative patterns.

Fig. 3: Sensitivity analysis of the proposed traffic pattern extraction method.

Next, the Fig. 4a shows good performance of the proposed hybrid Pathmox algorithm with extracted 14 + 7 patterns compared to fixed clustering approach, with 8 patterns, in terms of obtained GEH values, where the size of each point is proportional to the number of days within that pattern. Further, performance of the methods is evaluated by the histogram of integrated variability bands, as presented in Fig. 4b, to overcome the limitation of GEH measure that strongly penalizes deviations in low-flow scenarios. Fig. 4b clearly indicates better performance of the hybrid Pathmox algorithm due to its property to reduce the variability in extracted patterns instead of optimizing GEH.



(a) GEH (b) Histogram

Fig. 4: Comparison of extracted traffic flow patterns at h|47 node by proposed and user pre-defined methods.

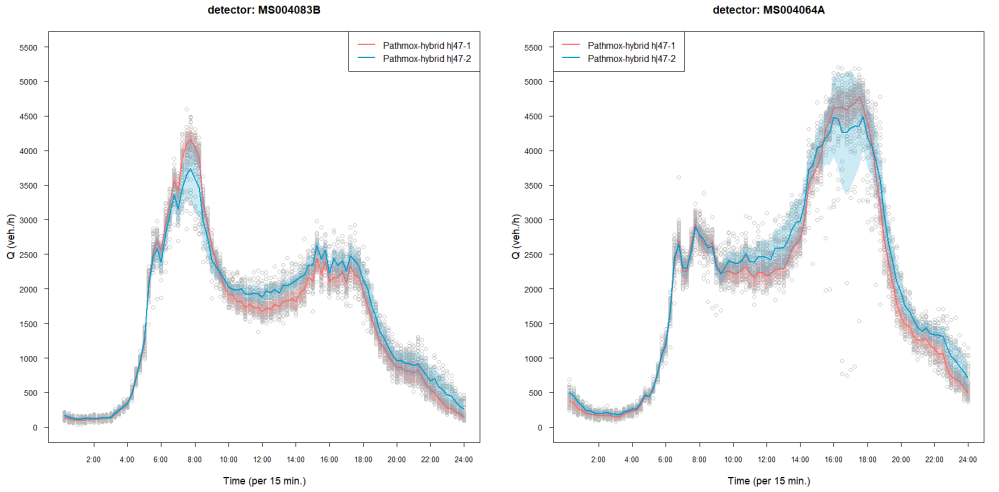
Fig. 5 shows how the proposed clustering method in phase 2 is able to efficiently capture trends in the data with almost non-overlapping conditional variability bands, especially in AM / PM peak hour intervals, which are characterized by sudden traffic flow increase and/or drop very difficult to predict, and which makes the contextual partitions to be very convenient after smooth low-flow night period. Additionally, variability bands in Fig. 5 reflect the efficient breakup of recurrent variability in the traffic flow. To fully support visual conclusions, Table 1 summarizes the GEH and variability results for two detectors and all detectors in the node h|47 obtained for the proposed as well as spatial clustering method based on *k*-means.

Table 1: Comparison of Pathmox-based 2 phase proposed method with spatial clustering in the node h|47.

	MS004083B		MS004064A		All	
Method	Spatial	Pathmox-hybrid	Spatial	Pathmox-hybrid	Spatial	Pathmox-hybrid
GEH (%)	92.1	**93.5**	88.5	**91**	90.8	**91.3**
Variability	2.5	**2.3**	3.5	**3.3**	2.79	**2.76**

5. Conclusions

This paper has discussed the potential of the use of hybrid Pathmox algorithm to extract the traffic flow profiles consistent among all the detectors in the network as a patterns. In determining performance of the proposed method, two benchmark studies were implemented to evaluate the extracted traffic flow patterns and their robustness. The contributions of this study include the following. The application of hybrid Pathmox algorithm to a set of traffic flow observations over a long time period on a freeway corridors provides evidence of the applicability of hybrid Pathmox algorithm to capture meaningful and robust patterns in a three-based configuration. The advantage of the hybrid approach to extract traffic patterns lies in the integration of the expert knowledge, but it is more flexible than user pre-defined clustering methods, as it incorporates variable interactions and is statistically supported process. Exploration



(a) Detector id MS004083B (AM peak high demand)

(b) Detector id MS004064A (PM peak high demand)

Fig. 5: Assessment of the proposed iterative clustering (red and cyan lines with symmetric variability bands at 95%) for the node h|47 in second scenario.

of the hybrid approach will overcome important weakness in spatial clustering and benefit early AM peak hour traffic prediction by using qualitative knowledge, when selection of early AM peak hour becomes hard after the night period when all patterns are similar. The authors suggest further generalization of the traffic pattern extraction method by incorporating variable selection and factor-level grouping in the contextual phase as well as improvement of the non-contextual iterative clustering part by introducing meta-heuristics algorithms. Also, in order to provide more robust pattern extraction, incidents will be removed using an Automatic Incident Detection module.

References

Attanasi, A., Meschini, L., Pezzulla, M., Fusco, G., Gentile, G., Isaenko, N., 2017. A hybrid method for real-time short-term predictions of traffic flows in urban areas. 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) , 878–883.

Bowman, A., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations. Oxford Statistical Science Series, OUP Oxford.

Breiman, L., Friedman, J., Stone, C., Olshen, R., 1984. Classification and Regression Trees. The Wadsworth and Brooks-Cole statistics-probability series, Taylor & Francis.

Feldman, O., 2012. The geh measure and quality of the highway assignment models, in: European Transport Conference 2012. Association for European Transport (AET) Transportation Research Board.

Gasser, T., Sroka, L., Jennen-Steinmetz, C., 1986. Residual variance and residual pattern in nonlinear regression. Biometrika 73, 625–633.

Han, Y., Moutarde, F., 2013. Statistical traffic state analysis in large-scale transportation networks using locality-preserving non-negative matrix factorisation. IET Intelligent Transport Systems 7, 283–295(12).

Han, Y., Moutarde, F., 2016. Analysis of large-scale traffic dynamics in an urban transportation network using non-negative tensor factorization. International Journal of Intelligent Transportation Systems Research 14, 36–49.

Lamberti, G., Aluja, T.B., Sanchez, G., 2016. The pathmox approach for pls path modeling segmentation. Applied Stochastic Models in Business and Industry 32, 453–468.

Salamanis, A., Margaritis, G., Kehagias, D.D., Matzoulas, G., Tzovaras, D., 2017. Identifying patterns under both normal and abnormal traffic conditions for short-term traffic prediction. Transportation Research Procedia 22, 665 – 674. 19th EURO Working Group on Transportation Meeting, EWGT2016, 5-7 September 2016, Istanbul, Turkey.

Soriguera, F., 2012. Deriving traffic flow patterns from historical data. Journal of Transportation Engineering 138, 1430–1441.

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometrics and Intelligent Laboratory Systems 2, 37 – 52. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

Yang, S., Wu, J., Qi, G., Tian, K., 2017. Analysis of traffic state variation patterns for urban road network based on spectral clustering. Advances in Mechanical Engineering 9.