Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 23 (2017) 901–921

Transportation Research Procedia

www.elsevier.com/locate/procedia

22nd International Symposium on Transportation and Traffic Theory

A Big Data Approach for Clustering and Calibration of Link Fundamental Diagrams for Large-Scale Network Simulation Applications

Ziyuan Gu[a], Meead Saberi[a,*], Majid Sarvi[b], Zhiyuan Liu[c]

[a] *Institute of Transport Studies, Department of Civil Engineering, Monash University, VIC 3800, Australia*
[b] *Department of Infrastructure Engineering, University of Melbourne, VIC 3010, Australia*
[c] *Jiangsu Key Laboratory of Urban ITS, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, School of Transportation, Southeast University, Nanjing 210096, China*

Abstract

Existing methods for calibrating link fundamental diagrams (FDs) often focus on a limited number of links and use grouping strategies that are largely dependent on roadway physical attributes alone. In this study, we propose a big data-driven two-stage clustering framework to calibrate link FDs for freeway networks. The first stage captures, under normal traffic state, the variations of link FDs over multiple days based on which links are clustered in the second stage. Two methods, i.e. the standard k-means algorithm combined with hierarchical clustering and a modified hierarchical clustering based on the Fréchet distance, are applied in the first stage to obtain the FD parameter matrix for each link. The calibrated matrices are input into the second stage where the modified hierarchical clustering is re-employed as a static approach resulting in multiple clusters of links. To further consider the variations of link FDs, the static approach is extended by modifying the similarity measure through the principle component analysis (PCA). The resulting multivariate time-series clustering models the distributions of the FD parameters as a dynamic approach. The proposed framework is applied on the Melbourne freeway network using one-year worth of loop detector data. Results have shown that (a) similar roadway physical attributes do not necessarily result in similar link FDs, (b) the connectivity-based approach performs better in clustering link FDs as compared with the centroid-based approach, and (c) the proposed framework helps achieving a better understanding of the spatial distribution of links with similar FDs and the associated variations and distributions of the FD parameters.

Peer-review under responsibility of the scientific committee of the 22nd International Symposium on Transportation and Traffic Theory.

Keywords: Link Fundamental Diagram; Calibration; Big Traffic Data; Clustering; Fréchet Distance; Traffic Dynamics

* Corresponding author. Tel.: +61 3 9905 0236
E-mail address: mead.saberi@monash.edu

Peer-review under responsibility of the scientific committee of the 22nd International Symposium on Transportation and Traffic Theory.
10.1016/j.trpro.2017.05.050

1. Introduction

Urban traffic management typically requires an understanding of traffic dynamics at the network level (Zheng et al., 2016), leading to extensive investigation into characterizing spatial and temporal travel patterns for network-level traffic flow analysis. To capture the spatial-temporal traffic dynamics in large-scale networks, simulation-based dynamic traffic assignment (DTA) models are deployed which require, however, accurately calibrated demand and supply inputs. As one of the major supply inputs, link fundamental diagrams (FDs) relate primary traffic flow variables with one another. Fig. 1 gives an example of off-line calibration for the dual-regime modified Greenshields traffic flow model (Mahmassani et al., 2009) using loop detector data from a freeway section on Western Ring Road, Melbourne in April 2011. Since transportation network models are critical in understanding network-wide traffic flow dynamics over time and space, calibration of link FDs needs to further consider the spatial-temporal features of traffic flow for achieving a better simulation performance. Also with the growing availability of big traffic data from mobile and infrastructure-based sources, a unique opportunity exists to improve the existing calibration and validation methods (Ozbay et al., 2014). Despite the rapid growth in the size and number of traffic data sets, traditional traffic data management tools and mining algorithms have not been sufficiently exploited. Limited research progress has been made both theoretically (Fahad et al., 2014; Zerhari et al., 2015) and empirically (Mudigonda and Ozbay, 2014; Ozbay et al., 2014) to provide valuable insights into clustering big data for different applications.

The majority of the existing research on FD calibration focuses on a limited number of links and involves grouping strategies assuming that similar roadway physical attributes lead to similar link FDs (refer to Section 1.1 for a comprehensive literature review). A few recent studies have explored a clustering-based framework for calibrating link FDs at either the section level (Jiang and Huang, 2009; Jiang et al., 2012) or the network level (Gu et al., 2016a, b). Nevertheless, as a supplementary approach to the traditional calibration methods, the big data-driven perspective has not been fully investigated at the network level particularly with regard to the spatial distribution of links with similar FDs and the associated variations and distributions of the FD parameters. Hence this study aims to address this concern and to extend the knowledge on FD calibration methods for freeway networks (refer to Section 1.2 for details of objectives and contributions).

1.1. Related literature

A vast body of literature has been devoted to the off-line calibration of link FDs for network simulation applications. Previous studies mainly focused on curve fitting using field data from a few number of days, either in normal situations (Del Castillo and Benitez, 1995; Smith et al., 1996; Leclercq, 2005) or under adverse weather conditions (Mahmassani et al., 2012; Hou et al., 2013). As a widely used technique for regression analysis, the least squares method (LSM) was typically employed to solve the curve fitting problem for dual- or multiple-regime traffic flow models (Dervisoglu et al., 2009; Li and Zhang, 2011). Because the LSM does not control for the sample selection bias, Qu et al. (2015) proposed a weighted least squares method (WLSM) for calibration of single-regime traffic flow models that better represents the congested regime. To improve network simulation applications, Chiu et al. (2010) further introduced the speed influencing region (SIR) for calibration of an anisotropic mesoscopic simulation (AMS) model.

We identify two limitations in these studies: (a) a steady state analysis of aggregated traffic data was applied without considering traffic flow dynamics, and (b) link FDs were assumed deterministic rather than stochastic. To overcome the first limitation, Zhong et al. (2015) proposed an automatic calibration method where a bi-level optimization problem was formulated to address the issue of data variability. The upper level aimed to minimize a merit function, i.e. the discrepancy between simulated and observed data, while the lower level was a cell transmission model (CTM). The studied freeway section was separated into different cells for simulation and calibration rather than treated as a whole, suggesting that spatial heterogeneity of link FDs was considered within the proposed method. Similarly, a probabilistic graphical model (Muralidharan et al., 2011) was proposed to reflect the probabilistic distribution of the FD parameters in simulation, i.e. the FD parameters were considered variables rather than constants. To address the

second concern, Wang et al. (2013) proposed a stochastic speed-density relationship as an extension to the well-established deterministic one where a term representing randomness was added to reflect the variability of traffic flow characteristics. Nevertheless, despite several extensions made to the traditional calibration methods, little progress has been achieved for calibration of link FDs at the network level.

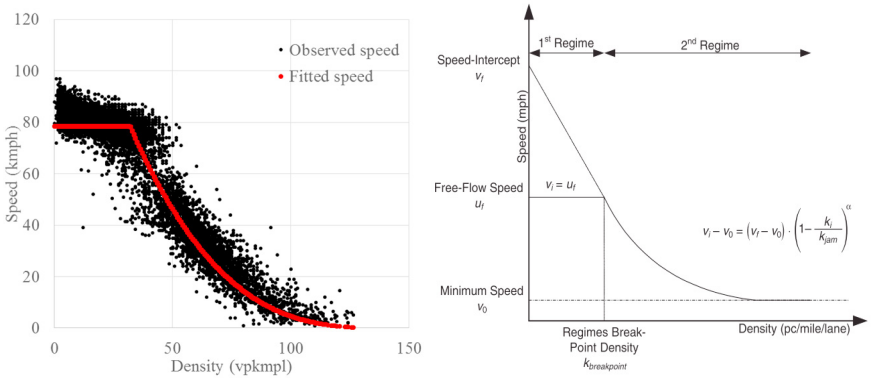


Fig. 1. Off-line calibration for a freeway section on Western Ring Road, Melbourne in April 2011: (a) fitted speed vs. observed speed; (b) the dual-regime modified Greenshields traffic flow model. Source: Hou et al. (2013).

Network-wide analysis of traffic data using different types of clustering techniques is well established in the literature. Clustering traffic data (a) helps to achieve a better understanding of the spatial distribution of traffic flow characteristics and provides traffic zone division with auxiliary decision support (Hu et al., 2011), (b) facilitates the representation of the network with limited storage in terms of traffic flow patterns (Banaei-Kashani et al., 2011), and (c) assists with time-dependent planning processes such as finding shortest paths (Kanoulas et al., 2006). The majority of the relevant literature, however, focused on using clustering techniques to categorize and analyze traffic conditions, either for flow pattern recognition (Weijermars and Van Berkum, 2005; Zheng et al., 2008; Banaei-Kashani et al., 2011; Hu et al., 2011; Kim and Mahmassani, 2015; Mudigonda and Ozbay, 2015; Saeedmanesh and Geroliminis, 2016) or for traffic state identification/prediction (Stutz and Runkler, 2002; Xia and Chen, 2007; Azimi and Zhang, 2010; Xia et al., 2012; Celikoglu and Silgu, 2016). Though a few studies have been inspired recently where clustering techniques were incorporated into the calibration procedure (Sun and Zhou, 2005; Jiang and Huang, 2009; Jiang et al., 2012), the main purpose of clustering was to determine the breakdown points of traffic flow and hence, the methods themselves were only suited for section-level calibration. Furthermore, despite numerous studies applying clustering techniques to investigate the spatial distribution of traffic flow patterns, the majority used either speed or flow profiles alone as the major inputs. Limited research considers clustering link FDs where traffic flow bi-variables are utilized.

1.2. Objectives and contributions

The main objectives of this study are twofold: (a) to develop a big-data driven approach for clustering and calibration of link FDs for freeway networks, and (b) to achieve a better understanding of the spatial distribution of links with similar FDs and the distributions and variations of the FD parameters. A few limitations of the existing calibration methods are identified: (a) the majority of the methods focus on specific freeway sections rather than the network as a whole, (b) when such methods are extended for network-level analysis, grouping strategies (Chiu et al., 2011) are typically employed that are largely dependent on roadway physical attributes alone, and (c) field data from only a short time period (e.g. a single day) are used resulting in the issue of "typical day" (Ozbay et al., 2014). It is a common practice to calibrate, using traffic data observed during a small time period, FDs for links with similar roadway physical attributes (e.g. mainline/weaving section/ramp by number of lanes and speed limit) without further internal differentiation, which may be considered a type of supervised machine learning unable to discover features in data by its own (Mohri et al., 2012). This limitation may lead to inaccurate calibration results and hence, motivates

the need to explore more advanced big data-driven unsupervised machine learning where link FDs are clustered based on the observed traffic flow characteristics.

To the best of our knowledge, few studies have considered clustering techniques to further differentiate between links with similar roadway physical attributes but with different FDs. An agglomerative hierarchical clustering based on k-means was proposed to calibrate speed-density relationships in a mesoscopic traffic simulator (Jiang and Huang, 2009; Jiang et al., 2012). The method mainly focused on separating observed traffic data (training data sets) and using the locally weighted regression (LWR) for speed estimation. Saeedmanesh and Geroliminis (2016) recently developed a three-step clustering algorithm to investigate link flow patterns at the network level, but the main contribution was on network partitioning using link density or speed for perimeter control based on macroscopic fundamental diagram (MFD). The framework by Gu et al. (2016a, 2016b) sheds some light into the introduced problem but was constrained to steady state analysis. Therefore, this study further develops the calibration methods for link FDs from a network perspective, which is among the very first of its kind that will provide valuable insights to supplement and extend the traditional calibration methods.

Building upon the existing studies, a two-stage clustering framework for calibrating link FDs for freeway networks is proposed. To capture the variations of link FDs over multiple days under normal traffic state, two methods are applied in the first stage using big traffic data from hundreds of sensors over multiple months across a large-scale freeway network, one being the standard k-means algorithm combined with hierarchical clustering and the other being a modified hierarchical clustering based on the Fréchet distance. The modified hierarchical clustering is re-employed in the second stage as a static approach to cluster links with similar FDs. A multivariate time-series clustering is also proposed as a dynamic approach by further taking into account the variations of link FDs. Therefore, the main contributions of this study include:

- A two-stage clustering framework is proposed for calibrating link FDs at the network level. The genericity of the methodology enables calibration for different combinations of link FDs and road types.
- The framework is big data- and network-driven which overcomes the “typical day” problem of the traditional calibration methods. We demonstrate that similar roadway physical attributes do not necessarily result in similar link FDs (yet an assumption of grouping strategies in many large-scale simulation applications).
- By taking into account the variations of link FDs over multiple days and months, the proposed method is able to characterize the spatial distribution of links with similar FDs and to model the associated variations and distributions of FD parameters.

The remainder of this paper is organized as follows. Section 2 presents the two-stage clustering framework. Section 3 analyzes the results and discusses the main findings. Section 4 concludes the paper.

2. Methodology: a two-stage clustering framework

Fig. 2 presents the two-stage clustering framework proposed in this study. The two parallel paths in both the first and second stages represent two alternative methods. Before the overall clustering procedure, data processing is needed and thus presented in Section 2.1. Section 2.2 discusses the two methods applied in the first stage with emphasis on the comparison between the centroid- and connectivity-based approaches. Section 2.3 presents and compares the two methods applied in the second stage.

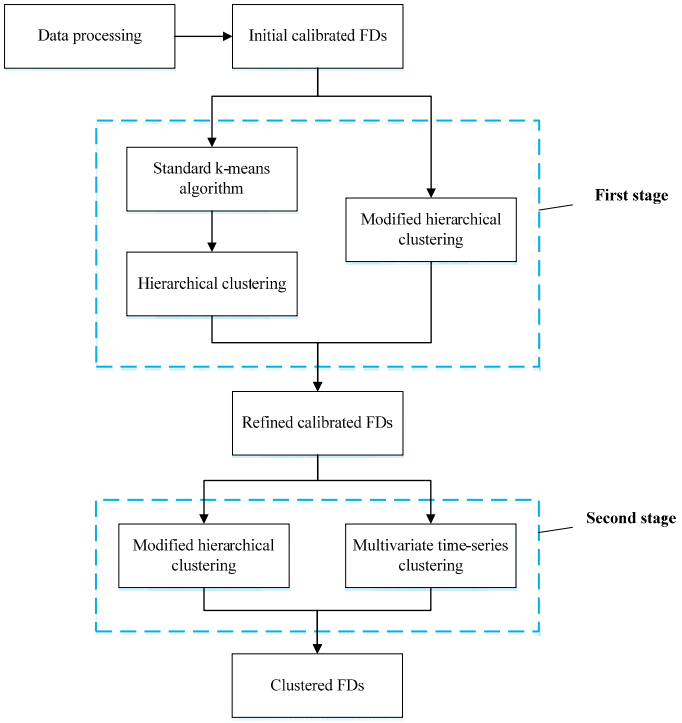


Fig. 2. A flowchart representation of the proposed two-stage clustering framework.

2.1. Data processing: parameter estimation

One-year worth of archived freeway loop detector data at one-minute intervals from Melbourne, Australia were used in this study. Two primary traffic flow variables were obtained, i.e. speed and occupancy. Before data processing, we filtered out observations with (a) negative speed or speed higher than 150 kmph, (b) negative occupancy or occupancy higher than 100%, and (c) speed lower than 30 kmph and occupancy lower than 10%. To calibrate link FDs against field data, occupancy (the percentage of time the detection zone of a detector is occupied by vehicles) needs to be converted to density (the number of vehicles occupying a given amount of roadway space). In line with Hou et al. (2013), the following relationship between the two variables was used as in Cassidy and Coifman (1997),

$$k = \frac{52.8}{L_v + L_s} \times o \tag{1}$$

where k is the density (vehicles per mile per lane/vpmpl); L_v and L_s are the average lengths for vehicles and sensors, assumed to be 5 and 2 m respectively (approximately 16.4 and 6.5 ft); o is the occupancy (%). A prerequisite for using Eq. (1), however, is to know the distribution of the detected vehicle lengths for each link across the network (Leclercq, 2005). Such information is not available through loop detectors and hence, an average vehicle length was assumed. Since insufficient data from the congested regime may lead to inaccurate calibration results, links with the majority of observations at the free-flow regime (i.e. the maximum density observed over multiple months not exceeding 50 vpmpl or equivalently 31 vpkmpl) were filtered out during data processing, resulting in 239 links for subsequent calibration and clustering analysis. Traffic data recorded during weekends and midnights (from 11 pm to 5 am) were also excluded.

Empirical evidence has shown that the dual-regime modified Greenshields traffic flow model can well represent freeway traffic for simulation purposes (Mahmassani et al., 2009) which was thus used in this study and mathematically expressed in Eq. (2),

$$v_i = \begin{cases} u_f & 0 < k_i < k_{bp} \\ v_0 + (v_f - v_0)\left(1 - k_i/k_j\right)^{\alpha} & k_{bp} < k_i < k_j \end{cases} \tag{2}$$

where v_i and k_i are the speed and the density on link i; u_{f}, v_0 and v_{f} are the free-flow speed, the jam speed and the intercept speed; k_{bp} and k_{j} are the breakpoint density and the jam density; α is a shape parameter. A few empirical studies (Chung et al., 2007; Dervisoglu et al., 2009; Saberi and Mahmassani, 2013; Kontorinaki et al., 2016) have revealed the capacity drop phenomenon when calibrating or simulating freeway traffic flow, i.e. $u_{\mathrm{f}}\left(k_{\mathrm{bp}}^{-}\right) > v_0 + (v_{\mathrm{f}} - v_0)\left(1 - k_{\mathrm{bp}}^{+}/k_{\mathrm{j}}\right)^{\alpha}$. Note that the capacity here refers to the maximum observed flow rate rather than the pre-breakdown flow rate as in Kim et al. (2010). The loss of the discharge flow is due to a pronounced increase in density. In this study, the link FDs were assumed to be continuous because non-consideration of the capacity drop helps reducing the number of the calibrated parameters and thus the size of the parameter matrix. Since the calibrated parameter matrices are the major inputs into the two-stage clustering framework (particularly for the multivariate time-series clustering), a smaller size leads to a better computational performance when dealing with big traffic data. It is noteworthy that non-consideration of the capacity drop remains a common problem for most existing large-scale network simulation applications which needs to be addressed as a future research direction. Nevertheless, the proposed framework as in Fig. 2 is able to consider the capacity drop (i.e. two-capacity FDs) and does not constrain itself to a specific type of link FD (see Section 3.1 for more details). Also with the fundamental equation of traffic flow, the speed-density relationship as investigated in this study can be readily transformed into either the flow-density or speed-flow relationships. Following the assumption, μ_{f} was replaced with $v_0 + (v_{\mathrm{f}} - v_0)\left(1 - k_{\mathrm{bp}}/k_{\mathrm{j}}\right)^{\alpha}$. k_{j} and v_0 were assumed 230 vpmpl (approximately 143 vpkmpl) and zero respectively rather than varying across different links. Hence the original Eq. (2) was transformed to Eq. (3),

$$v_i = \begin{cases} v_f\left(1 - k_{bp}/k_j\right)^{\alpha} & 0 < k_i < k_{bp} \\ v_f\left(1 - k_i/k_j\right)^{\alpha} & k_{bp} < k_i < k_j \end{cases} \tag{3}$$

where three parameters ($k_{\mathrm{bp}}, v_{\mathrm{f}}, \alpha$) jointly determine the shape of the link FD.

To calibrate link FDs against field data, curve fitting was performed on a daily basis for each link using the non-linear LSM. Chiabaut and Leclercq (2011) recently proposed an advanced calibration method based on the cumulative vehicle count (CVC) curves to estimate both the congestion wave speed and jam density. The method requires successive loop detectors without on- or off-ramps to ensure vehicle conservation. This requirement can hardly be met in this study. Given a set of N data points and a model function, i.e. $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ and $y = f(x, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ represents an n-dimensional vector of parameters to be calibrated, the LSM aims to minimize the sum of squared errors mathematically expressed in Eq. (4).

$$min\, SE = \sum_{i=1}^{N} \left(y_i - f(x_i, \boldsymbol{\theta})\right)^2 \tag{4}$$

Assuming that $f(x, \boldsymbol{\theta})$ is continuously differentiable, the first-order optimality condition of this unconstrained optimization problem implies that the minimum value of SE occurs when the gradient equals zero, i.e.

$$\frac{\partial SE}{\partial \theta_j} = -2 \sum_{i=1}^{N} \left(y_i - f(x_i, \boldsymbol{\theta})\right) \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_j} = 0 \tag{5}$$

for $\forall j \in [1, n]$. The obtained equation system is generally non-linear and can be solved numerically using the state-of-the-art algorithms (Björck, 1996; Qu et al., 2015), e.g. trust region methods. Because the dual-regime traffic flow

model was used, the LSM was employed instead of the WLSM since the sample selection bias was not present. A prerequisite for using the WLSM is that sufficient data are at the congested regime, otherwise the calculated weights may be highly inaccurate. During the calibration procedure, the initial calibrated FDs for each link were expressed as an $N \times 3$ matrix where N and 3 are the numbers of days and calibrated parameters respectively. Days with insufficient traffic data observed under the congested regime (i.e. the maximum density observed during the day does not exceed 50 vpmpl or equivalently 31 vpkmpl) were further excluded to ensure the accuracy of the calibration results, which implies that N is a variable that varies across different links rather than being a constant.

2.2. First stage: capturing the variations of link FDs over multiple days

Link FDs are influenced by external factors (e.g. weather, work zone or traffic management) and thus do not always reflect the normal daily traffic dynamics (Hou et al., 2013). Under the assumption that normal traffic state can be observed during most days of the year, two methods are applied to refine the initial calibration results and to extract those that reflect normal traffic dynamics.

2.2.1. The standard k-means algorithm combined with hierarchical clustering

The sequences of the calibrated parameters $(k_{\mathrm{bp}}, v_{\mathrm{f}}, \alpha)$ for each link can be treated as multiple observations in the three-dimensional Euclidean space. By applying a clustering technique, the cluster with the largest number of observations can be captured, considered a reflection of the variations of link FDs under normal traffic state. As one of the widely used clustering methods, the standard k-means algorithm (Lloyd, 1982) was employed (see Appendix A). In a centroid-based approach such as k-means, different clusters are represented by their respective central vectors. Each object is assigned to the closest cluster center. This status is updated through an iterative procedure till the termination criterion (usually the maximum number of iterations) is met. Because k-means is sensitive to the initial chosen centroids which may result in local minima, multiple randomly generated sets of starting centroids (10 in this study) were used where the best one was chosen. Two issues, however, were encountered when using the k-means method, one being the undesirable shape of the calibrated and clustered FDs and the other being the scatter of the fitted free-flow speeds within the extracted cluster.

The undesirable shape of the calibrated FDs results from the shape parameter α that lies within $(0,1]$. By calculating the second-order derivative of the congested regime in Eq. (2), Eq. (6) shows that, if α is estimated smaller than one, the term on the right-hand side $\alpha(\alpha-1)$ becomes negative resulting in a negative $\frac{d^2y}{dx^2}$ (note that $\frac{(v_{\mathrm{f}}-v_0)\left(1-k_i/k_{\mathrm{j}}\right)^{\alpha-2}}{{k_{\mathrm{j}}}^2}$ is positive).

$$\frac{d^2y}{dx^2} = \frac{\left(v_f - v_0\right)\left(1 - k_i/k_j\right)^{\alpha-2}}{{k_j}^2}\alpha(\alpha-1) \tag{6}$$

Hence the congested regime of FDs shows concavity rather than convexity. Also note that if α is estimated to be exactly one, the congested regime becomes a straight line connecting the points $\left(v_0 + (v_{\mathrm{f}} - v_0)\left(1 - k_{\mathrm{bp}}/k_{\mathrm{j}}\right)^{\alpha}, k_{\mathrm{bp}}\right)$ and $\left(v_0, k_{\mathrm{j}}\right)$. Despite the selection criteria used during data processing, observations for a number of links are scattered closely around k_{bp} with few in the high-density area, which results in a concave shape or straight line of the fitted congested regime. To address this issue, a threshold on α (i.e. $\alpha \geq 1.5$) was used for selection of the calibrated FDs.

The scatter of the fitted free-flow speeds within the extracted cluster may result from the applied clustering method per se. Due to the large similarity between the congested regimes, the k-means method tends to give less weight to the free-flow regimes (unadjusted weighting of multi-dimensional data) and hence, the calibrated FDs are clustered but with noticeable differences in the fitted free-flow speeds. To further improve the clustering performance, the method of interval selection can be incorporated.

Remark 1. The fitted free-flow speeds follow a Gaussian distribution.

Proof. A few studies (Donnell et al., 2009; Fazio et al., 2014) have shown that the observed free-flow speeds on freeways can be modeled as a Gaussian distribution,

$$V \sim N(\mu, \sigma^2) \tag{7}$$

where μ and σ^2 are unknown parameters to be estimated. Under the assumption that the set of the daily observed free-flow speeds $\{V_1, V_2, \dots, V_k\}$ is a random sample of size k drawn from the population, the sample mean $\overline{V}$ can also be modeled as a Gaussian distribution, expressed in Eq. (8).

$$\overline{V} \sim N(\mu, \sigma^2/k) \tag{8}$$

The non-linear LSM was used for calibration and hence, the following minimization problem was solved for the free-flow regime.

$$\min_{u_f} SE = \sum_{i=1}^{k} \left(V_i - u_f\right)^2 \tag{9}$$

By calculating the first-order derivative of Eq. (9),

$$\frac{dSE}{du_f} = -2\left(\sum_{i=1}^{k} V_i - k u_f\right) \tag{10}$$

the solution to the minimization problem can be obtained by setting $\frac{dSE}{du_f} = 0$, shown in Eq. (11).

$$u_f = \frac{1}{k}\sum_{i=1}^{n} V_i = \overline{V} \tag{11}$$

Therefore, the fitted free-flow speeds follow the same Gaussian distribution as in Eq. (8). By using the maximum likelihood estimation (MLE), the unknown parameters in Eq. (8) can be estimated.

$$\mu^* = \overline{u}_f \tag{12}$$

$$(\sigma^2/k)^* = \frac{1}{n}\sum_{i=1}^{n}\left(u_f^i - \overline{u}_f\right) \tag{13}$$

Based on Eqs. (12-13), a selection interval can be applied to refine the initial calibrated and clustered FDs, e.g. $\left(\mu^* \pm \sqrt{(\sigma^2/k)^*}\right)$.

Remark 2. The method of interval selection should satisfy the regularity condition that the sample size k is constant rather than varying across different days.

Since traffic data used in this study were not obtained from controlled experiments, we need to further filter the observations to satisfy Remark 2. Alternatively, the method of hierarchical clustering (see Appendix B) can be applied. It groups the objects into a binary hierarchical cluster tree based on a pre-defined similarity measure and cuts off the tree when further combinations lead to undesirable clusters for one of several reasons (see Leskovec et al. (2014) for more details). As a connectivity-based approach, the hierarchical clustering relates each object to its closest neighbors rather than using the notion of cluster center as in a centroid-based approach (e.g. k-means). In this study, the similarity measure used in the hierarchical clustering is the Manhattan distance. Its physical meaning in the context of clustering

link FDs is the absolute difference between the fitted free-flow speeds. The initial calibrated and clustered FDs where the fitted free-flow speeds exhibit a large discrepancy from the majority is excluded so that the maximum absolute difference is maintained no greater than five. To ensure that the FD variations are well-captured over multiple days, links are further discarded if the number of the calibrated FDs within the extracted cluster is less than 20.

2.2.2. A modified hierarchical clustering based on the Fréchet distance

As shown in Section 2.2.1, the scatter of the fitted free-flow speeds within the extracted cluster may result from the nature of k-means. Despite close proximity of a pair of objects (i.e. small Euclidean distance), the link FDs based on the two sets of the calibrated parameters may still present distinct curve patterns. Consider a simple function $y = ax^b$ with two parameters (α, β) and three points (1,1), (2,1), and (1,2). Though the Euclidean distances between (1,1) and the other two points are both one, the resulting functions with (1,1) and (2,1) are first-order polynomials whereas with (1,2), it represents a quadratic curve that is distinct from the other two. Nevertheless, the k-means method may still consider a single cluster for these three points leading to undesirable results.

Remark 3. The closeness between objects in the multi-dimensional Euclidean space does not necessarily lead to the similarity between curve patterns.

Since two types of clustering methods need to be integrated sequentially, the resulting combination method itself poses some degree of encumbrance for implementation. To address this concern, we further investigate the applicability of connectivity-based approaches and propose a modified hierarchical clustering based on the Fréchet distance. When clustering link FDs using big traffic data, the connectivity-based approach performs better than the centroid-based one for two reasons: (a) as per Remark 3, the centroid-based approach does not guarantee the similarity between curve patterns, and (b) the centroid-based approach typically applies an iterative procedure that significantly increases the computational burden for analyzing big traffic data.

The Fréchet distance, δ_{F}, is a measure of similarity between curves (Fréchet, 1906). A fundamental study on the computational properties of the Fréchet distance between two polygonal curves was first conducted by Alt and Godau (1995). Due to the computational complexity of the algorithm that involves the parametric search technique, Eiter and Mannila (1994) proposed a discrete variation termed the coupling distance δ_{dF}. The authors showed that the computational time is reduced from $O(pq \log^2 pq)$ to $O(pq)$ where p and q are the numbers of segments on the polygonal curves, and that δ_{dF} provides a good approximation to δ_{F}. A few recent studies have also been conducted either to improve the computational efficiency (Agarwal et al., 2014) or to consider the presence of outliers (De Carufel et al., 2014).

We define a curve as a continuous mapping $f: [a, b] \to V$ where $a, b \in \mathcal{R}$ and $a \leq b$. Given two curves $f: [a, b] \to V$ and $g: [a', b'] \to V$, the Fréchet distance is defined in Eq. (14),

$$\delta_F(f, g) = \inf_{\alpha,\beta} \max_{t\in[0,1]} d\left(f(\alpha(t)), g(\beta(t))\right) \tag{14}$$

where α, β are arbitrary non-decreasing continuous functions from [0,1] onto $[a, b]$ or $[a', b']$. To compute the Fréchet distance between arbitrary curves, the polygonal curves are typically introduced as an approximation. A polygonal curve is defined as $P: [0, n] \to V$ where n is a positive integer such that for $\forall i \in \{0,1, \dots, n-1\}$, $P(i + \lambda) = (1 - \lambda)P(i) + \lambda P(i + 1)$. Let P and Q be the polygonal curves and $\sigma(P) = (u_1, \dots, u_p)$ and $\sigma(Q) = (v_1, \dots, v_q)$ be the sequences of the endpoints of the piecewise functions. A coupling L between P and Q, i.e. $\{(u_{a_1}, v_{b_1}), (u_{a_2}, v_{b_2}), \dots, (u_{a_m}, v_{b_m})\}$, is a sequence of distinct pairs from $\sigma(P) \times \sigma(Q)$ such that $a_1 = b_1 = 1$, $a_m = p$, $b_m = q$, and for $\forall i \in \{1, \dots, m-1\}$, $a_{i+1} = a_i$ or $a_i + 1$ and $b_{i+1} = b_i$ or $b_i + 1$. The length of L, denoted by $\|L\|$, is the length of the longest link in L, i.e. $\|L\| = \max_{\forall i\in[1,m]} d(u_{a_i}, v_{b_i})$. Hence the discrete Fréchet distance between two polygonal curves P and Q is defined in Eq. (15).

$$\delta_{dF}(P,Q) = min\|L\| = min \max_{\forall i\in[1,m]} d\left(u_{a_i}, v_{b_i}\right) \tag{15}$$

The pseudo-code for computing $\delta_{\mathrm{dF}}(P,Q)$ using dynamic programming can be found in Gu et al. (2016b). By introducing $\delta_{\mathrm{dF}}(P,Q)$ into the first step of the hierarchical clustering presented in Appendix B, the modified hierarchical clustering is obtained. Note that the average $\delta_{\mathrm{dF}}(P,Q)$ between all pairs of objects in any two clusters is used rather than the nearest neighbor. To cut off the constructed binary hierarchical tree into multiple clusters, a cut-off criterion should be determined through trial-and-error. Similar to the combination method, links are further discarded if the number of the calibrated FDs within the extracted cluster is less than 20.

2.3. Second stage: clustering links with similar FDs

To cluster links with similar FDs obtained in the first stage, the modified hierarchical clustering is first re-employed in the second stage as a static approach. Since the calibrated FDs for each link are mathematically expressed as an $N \times 3$ matrix where each row represents a sequence of the calibrated parameters for a single day (an observation), the mean value of each calibrated parameter, i.e. the centroid of observations, is calculated upon which a representative link FD is built and input into the modified hierarchical clustering. Hence the static approach does not consider the variations of link FDs over multiple days which is well-suited for deterministic network simulations where link FDs are assumed constant. To further account for this variability that is not addressed by the static approach, a multivariate time-series clustering method (by further modifying the similarity measure used in the modified hierarchical clustering) is also proposed in the second stage as a dynamic approach based on Singhal and Seborg (2005). The authors developed a modified k-means method to analyze multivariate time-series data using the principle component analysis (PCA) similarity factor (Krzanowski, 1979) and the Mahalanobis distance similarity factor (Singhal and Seborg, 2002). To apply this method in the context of multi-day FD pattern recognition, the similarity measure used in the modified hierarchical clustering is reformulated consisting of two similarity factors, i.e. the PCA similarity factor (S_{PCA}) and the Fréchet distance similarity factor (S_{FD}). Since the parameter variations are considered, the dynamic approach is able to model the parameter distributions and thus well-suited for stochastic network simulations where link FDs are assumed varying over different days (the FD parameters vary according to the modeled distributions).

PCA is a commonly used multivariate statistical technique that calculates the principal directions of variability in the data (Jackson, 1991), i.e. the principle components (PCs). The PCA similarity factor quantifies the similarity measure between two data sets based on the comparison of their PCs. The number of PCs k is usually determined such that k PCs are able to represent at least 95% of the total variance in the data. The geometric interpretation of the PCA similarity factor is that it is the sum of squares of the cosines of angles between each pair of PCs. With a modification on the original PCA similarity factor to further consider the amount of variance explained by each PC (Johannesmeyer, 1999), the mathematical expression of S_{PCA} is finalized in Eq. (14),

$$S_{PCA}(L,M) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\left(\lambda_i^L \lambda_j^M\right) cos^2\,\theta_{ij}}{\sum_{i=1}^{k}\lambda_i^L \lambda_i^M} \tag{16}$$

where λ_i^L and λ_j^M are the i^{th} and j^{th} eigenvalues of PC subspaces of data sets L and M; θ_{ij} is the angle between the i^{th} and j^{th} PCs of L and M; k is the number of PCs.

Unlike the PCA similarity factor, the introduction of the Fréchet distance similarity factor compares two data sets that have the same spatial orientation but are located far apart (Singhal and Seborg, 2002). Therefore it serves as a supplement, particularly when two data sets have similar PCs but the numerical values of the variables are fairly different. Different from Singhal and Seborg (2005) where the Mahalanobis distance was used and incorporated into the Gaussian probability density function, here the Fréchet distance similarity factor relates the Fréchet distance with an empirical distribution calibrated with real data. To achieve the same order of magnitude as S_{PCA} (ranging from zero to one), the original Fréchet distance needs to be scaled down into S_{FD}. Consider a set of pairwise Fréchet distances

represented by a probability density function $\varphi(x)$ where x is the value of the distance metric, the Fréchet distance similarity factor is defined as the probability that the distance metric is not smaller than the computed value, i.e.

$$S_{FD}(P,Q) = \int_{\delta_{dF(P,Q)}}^{+\infty} \varphi(x)\,dx \tag{17}$$

subject to $S_{\mathrm{FD}}(P,Q) = \int_0^{+\infty} \varphi(x)\,dx = 1$. Essentially this resembles the cumulative distribution function of the pairwise Fréchet distance. Note that P and Q are the approximated polygonal curves of the calibrated FDs parameterized by the mean vectors of L and M respectively.

To combine S_{PCA} and S_{FD} into a single/composite similarity factor, a new distance metric SF is further introduced as a linear combination/weighted average of the two original similarity factors, expressed in Eq. (18),

$$SF = \alpha_1 S_{PCA} + \alpha_2 S_{FD} \tag{18}$$

where α_1 and α_2 are weightings subject to $\alpha_1 + \alpha_2 = 1$. By replacing $\delta_{\mathrm{dF}}(P,Q)$ with SF in the first step of the modified hierarchical clustering, a multivariate time-series clustering method is obtained. Sensitivity analysis has shown that $\alpha_1 = \alpha_2 = 0.5$ produces the best performance for this linear combination (Singhal and Seborg, 2005), which was thus used in this study.

Remark 4. The major difference between the modified hierarchical clustering and the multivariate time-series clustering is the used similarity measure, i.e. δ_{dF} vs. SF.

3. Results and discussion

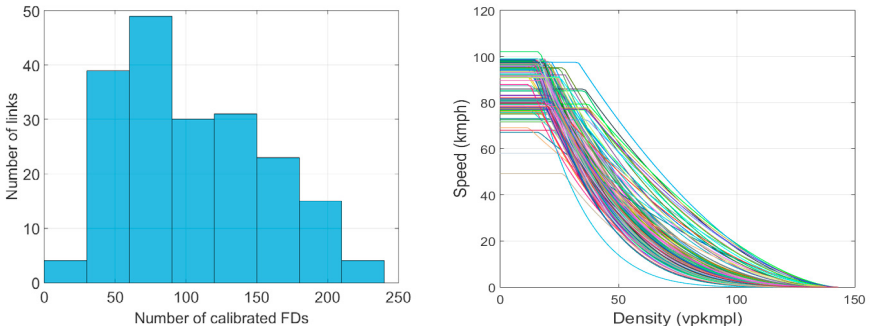
MySQL was used for big traffic data storage and processing. MATLAB was used to implement the two-stage clustering framework. The initial number of links was 319 which reduced to 239 after data cleaning and processing. After the implementation of the first stage, 44 links were further excluded resulting in a total of 195 links. These links were input into the second stage and the results are discussed below.

3.1. Static approach: clusters of links with representative FDs

With Eq. (3), 195 links were ultimately selected and calibrated, for each of which a matrix representing the variations of the calibrated parameters over multiple days was obtained. During the calibration procedure, the average values of the adjusted R^2 and the root mean square error (RMSE) were 0.81 and 3.83 respectively. Fig. 3(a) shows that for the majority of links, the number of the calibrated FDs ranges from 50 to 150. By calculating the mean value of each calibrated parameter, a representative FD was obtained for each link, shown in Fig. 3(b). Though these individual calibrated FDs exhibit, as expected, significant scatter, such scatter is largely smoothed once the links are clustered based on the calibrated parameters.

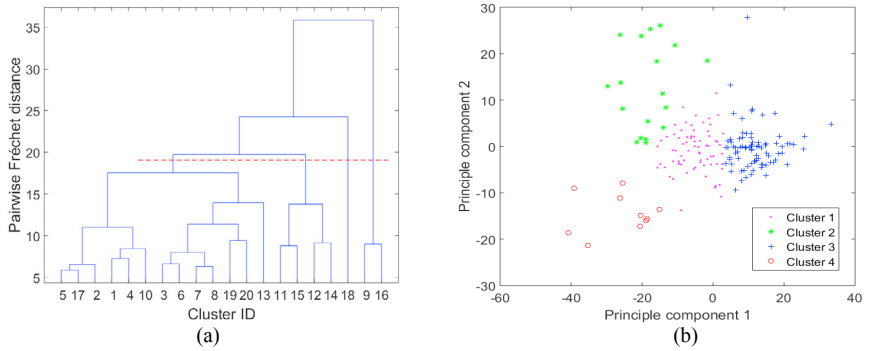
Fig. 4(a) shows the binary hierarchical tree resulting from the modified hierarchical clustering. Objects with the closest proximity (i.e. the link FDs with the least Fréchet distance) are connected and combined into a new object, represented by the multiple blue lines in the figure. Due to space limitation, only the top 20 hierarchical clusters are presented. To further illustrate the limitation of the k-means method, Fig. 4(b) applies the PCA technique to visualize the clustering results obtained from k-means in a two-dimensional Euclidean space. Despite four distinct clusters being identified, the points within the green- and red-colored clusters are far more scattered as compared with those within the other two, suggesting that the link FDs represented by these points may still exhibit noticeable shape differences. If the hierarchical tree is cut off where four clusters are formed (as shown by the red dash line in Fig. 4(a)), the maximum pairwise Fréchet distance can be as large as nearly 20, which also suggests that the link FDs within certain clusters may not exhibit similar patterns.

Fig. 5(a) shows that five clusters are obtained using a predetermined cut-off criterion (the Fréchet distance valued at five). Since the majority of clusters (23 out of 28) contain less than 10 links (and only one for most of them), the calibrated FDs within these clusters may exhibit considerable dissimilarity that are better treated individually. Despite only five clusters containing more than 10 links, over 70% (148 out of 195) of links are included. This observation suggests that a large proportion of links exhibit similar FD patterns. For each of the five clusters, a representative FD may serve as a reasonable approximation. Fig. 5(b) shows the five representative FDs (see Fig. 8(a) for the flow-density relationships) and Table 1 summarizes the numerical details.



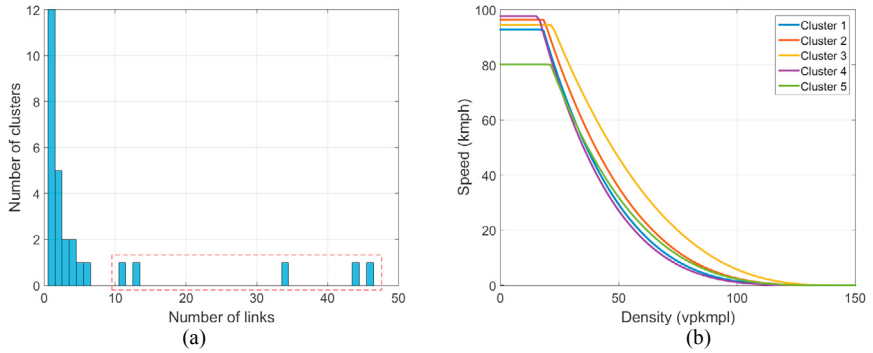
(a) (b)

Fig. 3. Calibration results from the static approach: (a) number of links vs. number of the calibrated FDs; (b) 195 representative FDs.



(a) (b)

Fig. 4. A comparison between the connectivity- and centroid-based approaches: (a) the binary hierarchical tree resulting from the modified hierarchical clustering; (b) four distinct clusters identified by the k-means method.



(a) (b)

Fig. 5. Clustering results from the static approach without capacity drop: (a) five clusters obtained; (b) five representative FDs.

Table 1. Calibrated parameters for each cluster without capacity drop.

	Breakpoint density (vpkmpl)	Free-flow speed (kmph)	Intercept speed (kmph)	Capacity (vph)	Alpha	Number of links
Cluster 1	18.03	92.42	156.92	1880.21	3.90	13
Cluster 2	18.45	95.97	154.53	2079.47	3.42	46
Cluster 3	21.85	94.10	147.74	2438.76	2.70	11
Cluster 4	16.11	97.29	159.96	1821.04	4.13	34
Cluster 5	21.16	79.82	138.55	1864.43	3.42	44

Since the largest pairwise Fréchet distance is approximately 60, the applied cut-off criterion may be considered a tight threshold. One can slightly release the cut-off criterion so that more links will be included in the identified clusters with, however, reduced similarity. Therefore, this is essentially a trade-off between quality and quantity. By using the Fréchet distance as the similarity measure, the majority of links can be clustered because of the similar FD patterns whereas a few show significant individual features. This observation suggests that both clustering and classification (grouping strategies) should not be employed in a holistic manner unless links with distinct FD patterns are identified.

When calibrated and clustered links are mapped to the real freeway network, links within each cluster do not necessarily exhibit the same roadway physical attributes, shown in Fig. 6. For each cluster, the number of lanes varies significantly suggesting little direct connection between this physical attribute and the similarity between link FDs. In terms of speed limit, a good relation can be observed because a uniform pattern prevails in each of the five clusters. Nevertheless, links with the same speed limit are still partitioned into multiple clusters probably due to the dissimilarity in the congested regime (starting from the breakpoint density). Further investigation may be needed to understand the potential factors leading to this phenomenon other than the two examined here.

Fig. 6. Distributions of roadway physical attributes within each cluster: (a) number of lanes; (b) speed limit.

To further demonstrate that the proposed framework as in Fig. 2 is able to consider the capacity drop (i.e. two-capacity FDs) and does not constrain itself to a specific type of link FD (as argued in both Sections 1.2 and 2.1), we relaxed the assumption of non-consideration of the capacity drop and used a two-capacity triangular FD as in Dervisoglu et al. (2009) during the implementation of the static approach. The corresponding speed-density relationship is mathematically expressed in Eq. (19),

$$v_i = \begin{cases} u_f & 0 < k_i < k_{bp} \\ w\left(1 - k_j/k_i\right) & k_{bp} < k_i < k_j \end{cases} \tag{19}$$

where v_i and k_i are the speed and the density on link i; u_{f} is the free-flow speed; k_{bp} and k_{j} are the breakpoint density and the jam density; w is the slope of the congested regime of the flow-density relationship (negative), i.e. the congestion wave speed. The results are shown in Fig. 7 and Table 2. Due to the consideration of the capacity drop

(and thus one more parameter added to the parameter matrix), the number of clusters reduces to four (128 links included), shown in Fig. 7(a). Though the resulting representative FDs as shown in Fig. 7(b) look similar to those in Fig. 5(b), considering the capacity drop results in a vertical gap between the speeds on either side of the breakpoint density (discontinuous FDs) (see Fig. 8(b) for the flow-density relationships). From Table 2 we can observe that the value of the capacity drop varies. The least occurs for Cluster 2 where the free-flow speed is the lowest. With the increase of the free-flow speed, the percentage of the capacity drop also rises as shown by the other three clusters, suggesting that a higher free-flow speed (and thus a higher speed limit) may result in a more significant capacity drop for freeway links.

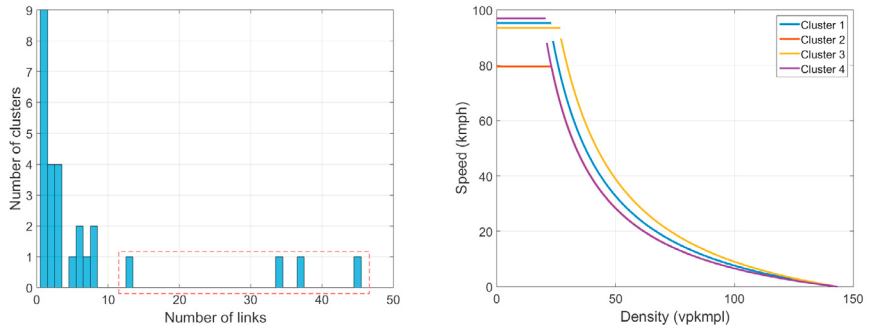


Fig. 7. Clustering results from the static approach considering the capacity drop: (a) four clusters obtained; (b) four representative FDs.

Table 2. Calibrated parameters for each cluster considering the capacity drop.

	Breakpoint density (vpkmpl)	Free-flow speed (kmph)	Capacity (vph)	Capacity drop (%)	Number of links
Cluster 1	23.33	95.28	2222.97	5.96	37
Cluster 2	23.44	79.54	1864.78	2.84	34
Cluster 3	26.75	93.51	2501.61	3.30	13
Cluster 4	20.78	96.99	2015.49	8.86	45

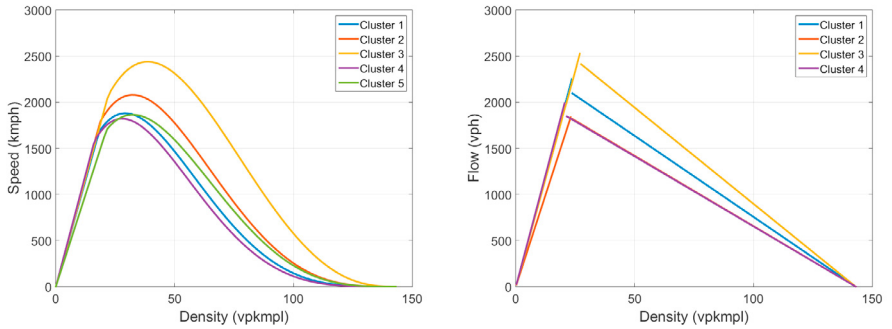


Fig. 8. The flow-density relationship for each cluster: (a) the dual-regime FDs without capacity drop; (b) the two-capacity triangular FDs considering the capacity drop.

3.2. Dynamic approach: clusters of links with distributions of the calibrated parameters

To further account for the variations of link FDs over multiple days, the inputs into the multivariate time-series clustering become, for each link, a set of varying link FDs rather than the invariant one assumed in the modified hierarchical clustering. The assumption of non-consideration of the capacity drop was used in the dynamic approach for simplicity, but can be easily relaxed as shown in the static approach in Section 3.1. An empirical distribution of the pairwise Fréchet distance needs to be modeled prior to the construction of the binary hierarchical tree. To specify

the distribution function, 17 well-known continuous distributions including the Gaussian, Weibull, logistic, etc. were tested against 18,915 pairwise Fréchet distances. The MLE was used to estimate the parameters of different distribution functions. The likelihood function as in Kim et al. (2010) is expressed in Eq. (20),

$$L = \prod_{i=1}^{n} f\left(\delta_{dF}^{i}\right)^{\theta_i} \left(1 - F\left(\delta_{dF}^{i}\right)\right)^{1-\theta_i} \tag{20}$$

where n is the number of observations; $\theta_i = 1$ if uncensored and 0 otherwise; $f(\cdot)$ and $F(\cdot)$ are the probability density function and the cumulative distribution function respectively. To select the best fitted distribution among the finite set of alternatives, the Bayesian information criterion (BIC) was used which is partially based on the likelihood function, defined in Eq. (21),

$$BIC = -2 \ln L_{max} + m \ln n \tag{21}$$

where $L_{\max}$ is the maximized value of the likelihood function; m is the number of the estimated parameters; n is the number of observations.

By calculating the BIC value for each tested distribution function and ranking these competing functions in an ascending order, the one with the lowest BIC value is chosen as the best candidate. Since the Weibull distribution provides the best fit as shown in Fig. 9, it was used and expressed in Eq. (22) in the form of the probability density function ($x \geq 0$). Note that the distribution applies to the pairwise Fréchet distances between the individual objects at the bottom of the hierarchical tree. During the construction of this hierarchical structure, the average distance metric between the newly-formed clusters was used rather than the farthest neighbor. As a result, the maximum value (less than 60 as shown in Fig. 9(a)) may decrease during the process (less than 40 as shown in Fig. 4(a)). Nevertheless, since the distribution pattern shown in Fig. 9 exhibits the unilateral long-tail feature (i.e. the cumulative probability reaches over 0.9 when the pairwise Fréchet distance is 30), we assume that the observed decrease is insignificant and hence, a fixed distribution function may serve as a reasonable approximation. The final similarity measure SF is expressed in Eq. (23). Note that the number of PCs is set three.

$$\varphi(x) = 0.099 \left(\frac{x}{17.57}\right)^{0.74} exp\left(-\left(\frac{x}{17.57}\right)^{1.74}\right) \tag{22}$$

$$SF = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3}\left(\lambda_i^L \lambda_j^M\right) cos^2\,\theta_{ij}}{2\sum_{i=1}^{3}\lambda_i^L \lambda_i^M} + \int_{\delta_{dF(P,Q)}}^{+\infty} \frac{\varphi(x)}{2} dx \tag{23}$$

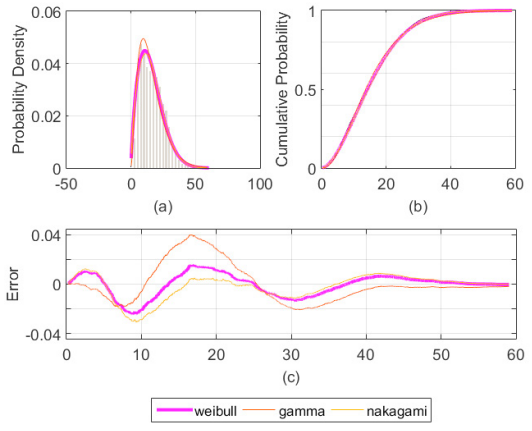


Fig. 9. The Weibull distribution of the pairwise Fréchet distance: (a) the probability density function; (b) the cumulative distribution function; (c) errors in the cumulative probability.

Similar to the static approach, the binary hierarchical tree resulting from the dynamic approach is shown in Fig. 10(a). Four clusters are obtained using a pre-determined cut-off criterion (SF valued at 0.08), shown in Fig. 10(b). A large proportion of links (nearly 70%) are included suggesting that the majority of links have similar FD variations whereas a few exhibit greatly differed variations. For each cluster, since the variations of link FDs are similar, the calibrated parameter matrix of each included link was combined resulting in a composite parameter matrix. This larger matrix is essentially a mixture that represents the similar FD variations of various links over different days, based on which the MLE in conjunction with the BIC was re-employed to model the parameter distributions. Fig. 11 shows that among all the tested distributions, the generalized extreme value (GEV) distribution, the lognormal distribution, and the gamma distribution can accurately reproduce for each cluster the distribution patterns of k_{bp}, v_{f}, and α respectively. Since the distribution of each FD parameter can be captured for each cluster by a distribution function with a specific set of parameters, a probabilistic link FD can be derived for each cluster where the FD parameters vary on a daily basis according to the modeled distribution. From a modeling perspective, the probabilistic link FD can be seen as a dynamic representation of the static link FD where the variations in traffic behavior are considered and captured (Muralidharan et al., 2011).

Fig. 12 shows the spatial distribution of the four clusters resulting from the dynamic approach. Links that do not experience much congestion are excluded. We can approximately represent these links with only the free-flow regime instead of calibrating a biased congested part of the FD. Nevertheless, the current framework is not able to distinguish between links represented with only the free-flow regime and with a complete FD. Eliminating this limitation is a direction for future research. A closer look at Fig. 12 reveals that links with similar FDs do exhibit some spatial correlations. Despite a few exceptions, links tend to locate in the same area provided that the calibrated FDs exhibit similar patterns. Though the speed limit may partially contribute to this observation as suggested by Fig. 6(b), further investigation into the land use pattern as well as the local driving environment may help better understand the causes.

The results from the dynamic approach may be scenario-based, i.e. dependent on the applied cut-off criterion. Hence a few more cut-off criteria were further investigated ranging from 0.1, 0.15 to 0.2. Results from the sensitivity analysis have shown that when the cut-off criterion is set no greater than 0.15, the number of the identified clusters remains unchanged (i.e. four). The GEV, lognormal, and gamma distributions can consistently represent the variations of the calibrated FD parameters. When the threshold is raised to 0.2, however, the performance of the dynamic approach drops significantly resulting in only three clusters with much greater dissimilarity. No distribution function is able to consistently model the variations of each calibrated FD parameter across all the clusters. Therefore, this value may serve as an upper bound for achieving a good performance of the dynamic approach.

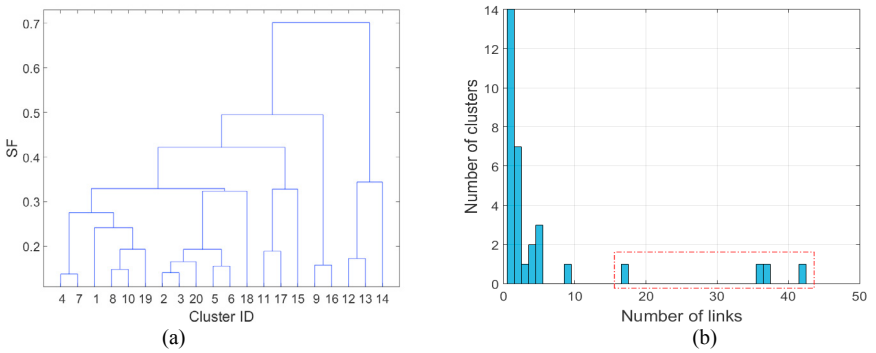


Fig. 10. Clustering results from the dynamic approach: (a) the binary hierarchical tree resulting from the multivariate time-series clustering; (b) four clusters obtained.

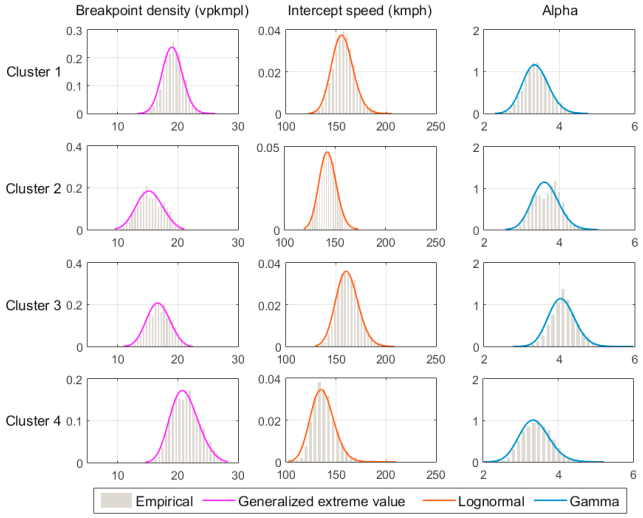


Fig. 11. Distributions of the calibrated FD parameters within each cluster resulting from the dynamic approach.

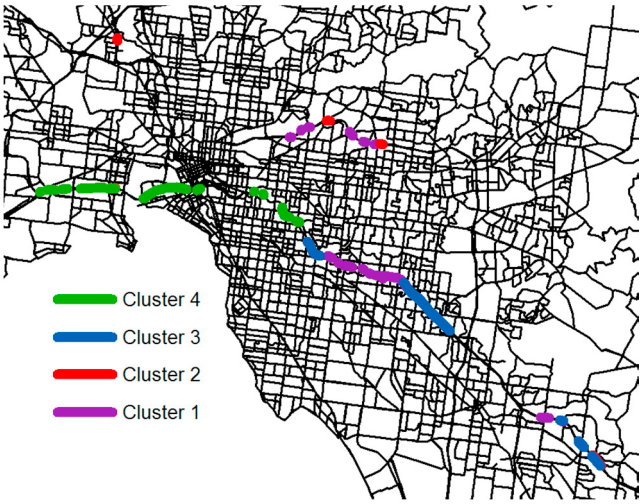


Fig. 12. The spatial distribution of the four clusters resulting from the dynamic approach.

4. Conclusion

Considering the variations in traffic behavior over multiple days and the increasing availability of big traffic data, a two-stage clustering framework is proposed in this study for calibrating link FDs for freeway networks. In the first stage, two parallel methods are presented and compared for capturing the variations of link FDs under normal traffic state. Due to the limitation of the k-means method, the hierarchical clustering is integrated sequentially resulting in a combination method. As a comparison, a connectivity-based approach is further proposed where the Fréchet distance similarity measure is incorporated into the hierarchical clustering. The resulting modified hierarchical clustering itself is able to control the dissimilarity between the fitted free-flow speeds within the extracted cluster. In the second stage, two parallel methods are proposed to cluster links with similar FDs, one being a static approach where link FDs are assumed invariant over multiple days and the other being a dynamic approach where a probabilistic link FD is derived for each cluster.

The proposed framework is applied to the Melbourne freeway network using one-year worth of one-minute aggregated data. Results have shown that links with similar FDs can be accurately identified and clustered. An important observation is that the majority of links exhibit similar FD patterns whereas a few show significant individual features. As a result, both clustering and classification (grouping strategies) should not be employed in a holistic manner unless links with distinct FD patterns are identified. The proposed framework is generic and can be applied regardless of the selected link FD shape and road type. Given any mathematical expression of the link FD and the required traffic data for calibration, the FD parameters can be obtained and input into the clustering procedure resulting in multiple clusters of links with similar FDs. Nevertheless, the results presented here are based on a specific freeway network and further investigation may be needed looking at different road networks. Note that in this study, each link is considered independent when calibrated. An interesting study by Muralidharan et al. (2011) assumed a joint distribution of FD parameters for a freeway stretch. Applying and extending this type of spatial correlation to the network level is difficult and remains an open question for future research. Also note that link FDs are assumed deterministic rather than stochastic, i.e. a one-to-one mapping relates density to speed (single-valuedness). As another direction for future research, the current framework may be further extended by assuming a one-to-many relationship (multi-valuedness) through a random term added to the deterministic FD **(**Wang et al., 2013**)**.

References

Agarwal, P.K., Avraham, R.B., Kaplan, H., Sharir, M., 2014. Computing the discrete Fréchet distance in subquadratic time. *SIAM Journal on Computing* 43(2), 429-449.

Alt, H., Godau, M., 1995. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications* 5(01n02), 75-91.

Azimi, M., Zhang, Y., 2010. Categorizing freeway flow conditions by using clustering methods. *Transportation Research Record: Journal of the Transportation Research Board*(2173), 105-114.

Banaei-Kashani, F., Shahabi, C., Pan, B., 2011. Discovering patterns in traffic sensor data, *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on GeoStreaming*. ACM, pp. 10-16.

Björck, A., 1996. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, PA.

Cassidy, M., Coifman, B., 1997. Relation among average speed, flow, and density and analogous relation between density and occupancy. *Transportation Research Record: Journal of the Transportation Research Board*(1591), 1-6.

Celikoglu, H.B., Silgu, M.A., 2016. Extension of Traffic Flow Pattern Dynamic Classification by a Macroscopic Model Using Multivariate Clustering. *Transportation Science*, 1-16.

Chiabaut, N., Leclercq, L., 2011. Wave velocity estimation through automatic analysis of cumulative vehicle count curves. *Transportation Research Record: Journal of the Transportation Research Board*(2249), 1-6.

Chiu, Y.-C., Bottom, J., Mahut, M., Paz, A., Balakrishna, R., Waller, T., Hicks, J., 2011. *Dynamic Traffic Assignment: A Primer*. Transportation Research Circular E-C153, Transportation Research Board, Washington, DC.

Chiu, Y.-C., Zhou, L., Song, H., 2010. Development and calibration of the Anisotropic Mesoscopic Simulation model for uninterrupted flow facilities. *Transportation Research Part B: Methodological* 44(1), 152-174.

Chung, K., Rudjanakanoknad, J., Cassidy, M.J., 2007. Relation between traffic density and capacity drop at three freeway bottlenecks. *Transportation Research Part B: Methodological* 41(1), 82-95.

De Carufel, J.-L., Gheibi, A., Maheshwari, A., Sack, J.-R., Scheffer, C., 2014. Similarity of polygonal curves in the presence of outliers. *Computational Geometry* 47(5), 625-641.

Del Castillo, J., Benitez, F., 1995. On the functional form of the speed-density relationship—II: empirical investigation. *Transportation Research Part B: Methodological* 29(5), 391-406.

Dervisoglu, G., Gomes, G., Kwon, J., Horowitz, R., Varaiya, P., 2009. Automatic calibration of the fundamental diagram and empirical observations on capacity, *Transportation Research Board 88th Annual Meeting*, Washington, DC.

Donnell, E., Hines, S., Mahoney, K., Porter, R., McGee, H., 2009. Speed Concepts: Informational Guide. FHWA-SA-10-001, US Department of Transportation, FHWA, Washington, DC.

Eiter, T., Mannila, H., 1994. Computing discrete Fréchet distance. Technical Report CD-TR 94/64, Information Systems Department, Technical University of Vienna, Vienna, Austria.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., Foufou, S., Bouras, A., 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing* 2, 267-279.

Fazio, J., Wiesner, B.N., Deardoff, M.D., 2014. Estimation of free-flow speed. *KSCE Journal of Civil Engineering* 18(2), 646-650.

Fréchet, M.M., 1906. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo* 22, 1-74.

Gu, Z., Saberi, M., Sarvi, M., Liu, Z., 2016a. Calibration and clustering of traffic flow fundamental diagrams for network simulation applications: a case study in Melbourne, *Proceedings of the 23rd World Congress on Intelligent Transport Systems*, Melbourne, Australia.

Gu, Z., Saberi, M., Sarvi, M., Liu, Z., 2016b. Calibration of Traffic Flow Fundamental Diagrams for Network Simulation Applications: A Two-Stage Clustering Approach, *Proceedings of the 19th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, Rio de Janeiro, Brazil.

Hou, T., Mahmassani, H., Alfelor, R., Kim, J., Saberi, M., 2013. Calibration of Traffic Flow Models under Adverse Weather and Application in Mesoscopic Network Simulation. *Transportation Research Record: Journal of the Transportation Research Board*(2391), 92-104.

Hu, C., Luo, N., Yan, X., Shi, W., 2011. Traffic flow data mining and evaluation based on fuzzy clustering techniques. *International Journal of Fuzzy Systems* 13(4), 344-349.

Jackson, J.E., 1991. *A User's Guide to Principal Components*. John Wiley & Sons, New York, US.

Jiang, Z., Huang, Y.-X., 2009. Parametric calibration of speed–density relationships in mesoscopic traffic simulator with data mining. *Information Sciences* 179(12), 2002-2013.

Jiang, Z., Shubin, L., Xiaoqing, L., 2012. Parameters Calibration of Traffic Simulation Model Based on Data Mining. *Journal of Transportation Systems Engineering and Information Technology* 12(6), 28-33.

Johannesmeyer, M.C., 1999. Abnormal situation analysis using pattern recognition techniques and historical data (master's thesis). University of California, Santa Barbara, CA.

Kanoulas, E., Yang, D., Tian, X., Donghui, Z., 2006. Finding Fastest Paths on A Road Network with Speed Patterns, *Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)*, pp. 10-10.

Kim, J., Mahmassani, H., Dong, J., 2010. Likelihood and duration of flow breakdown: modeling the effect of weather. *Transportation Research Record: Journal of the Transportation Research Board*(2188), 19-28.

Kim, J., Mahmassani, H.S., 2015. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. *Transp. Res. Part C* 59, 375-390.

Kontorinaki, M., Spiliopoulou, A., Roncoli, C., Papageorgiou, M., 2016. Capacity Drop in First-Order Traffic Flow Models: Overview and Real-Data Validation, *Transportation Research Board 95th Annual Meeting*, Washington, DC.

Krzanowski, W., 1979. Between-groups comparison of principal components. *Journal of the American Statistical Association* 74(367), 703-707.

Leclercq, L., 2005. Calibration of Flow-Density Relationships on Urban Streets. *Transportation Research Record: Journal of the Transportation Research Board*(1934), 226-234.

Leskovec, J., Rajaraman, A., Ullman, J.D., 2014. *Mining of Massive Datasets*. Cambridge University Press, Cambridge, UK.

Li, J., Zhang, H., 2011. Fundamental diagram of traffic flow: new identification scheme and further evidence from empirical data. *Transportation Research Record: Journal of the Transportation Research Board*(2260), 50-59.

Lloyd, S.P., 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28(2), 129-137.

Mahmassani, H., Dong, J., Kim, J., Chen, R.B., Park, B., 2009. Incorporating Weather Impacts in Traffic Estimation and Prediction Systems. FHWA-JPO-09-065, US Department of Transportation, FHWA, Washington, DC.

Mahmassani, H., Kim, J., Hou, T., Zockaie, A., Saberi, M., Jiang, L., Verbas, O., Cheng, S., Chen, Y., Haas, R., 2012. Implementation and Evaluation of Weather Responsive Traffic Estimation and Prediction System. FHWA-JPO-12-055, US Department of Transportation, FHWA, Washington, DC.

Mohri, M., Rostamizadeh, A., Talwalkar, A., 2012. *Foundations of Machine Learning*. MIT Press, Cambridge, MA.

Mudigonda, S., Ozbay, K., 2014. Using big data and efficient methods to capture stochasticity for calibration of macroscopic traffic simulation models *Symposium Celebrating 50 Years of Traffic Flow Theory*, Portland, Oregon.

Mudigonda, S., Ozbay, K., 2015. Robust calibration of macroscopic traffic simulation models using stochastic collocation. *Transp. Res. Part C* 59, 358-374.

Muralidharan, A., Dervisoglu, G., Horowitz, R., 2011. Probabilistic graphical models of fundamental diagram parameters for simulations of freeway traffic. *Transportation Research Record: Journal of the Transportation Research Board*(2249), 78-85.

Ozbay, K., Yang, H., Morgul, E.F., Mudigonda, S., Bartin, B., 2014. Big data and the calibration and validation of traffic simulation models, *Traffic and Transportation Simulation - Looking Back and Looking Ahead: Celebrating 50 Years of Traffic Flow Theory, a Workshop*. Transportation Research Board, Washington, DC, pp. 92–122.

Qu, X., Wang, S., Zhang, J., 2015. On the fundamental diagram for freeway traffic: A novel calibration approach for single-regime models. *Transportation Research Part B: Methodological* 73, 91-102.

Saberi, M., Mahmassani, H., 2013. Hysteresis and capacity drop phenomena in freeway networks: empirical characterization and interpretation. *Transportation Research Record: Journal of the Transportation Research Board*(2391), 44-55.

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

Singhal, A., Seborg, D.E., 2002. Pattern matching in multivariate time series databases using a moving-window approach. *Industrial & Engineering Chemistry Research* 41(16), 3822-3838.

Singhal, A., Seborg, D.E., 2005. Clustering multivariate time-series data. *Journal of Chemometrics* 19(8), 427-438.

Smith, W.S., Hall, F.L., Montgomery, F.O., 1996. Comparing the speed-flow relationship for motorways with new data from the M6. *Transportation Research Part A: Policy and Practice* 30(2), 89-101.

Stutz, C., Runkler, T.A., 2002. Classification and prediction of road traffic using application-specific fuzzy clustering. *IEEE Transactions on Fuzzy Systems* 10(3), 297-308.

Sun, L., Zhou, J., 2005. Development of multiregime speed-density relationships by cluster analysis. *Transportation Research Record: Journal of the Transportation Research Board*(1934), 64-71.

Wang, H., Ni, D., Chen, Q.Y., Li, J., 2013. Stochastic modeling of the equilibrium speed–density relationship. *Journal of Advanced Transportation* 47(1), 126-150.

Weijermars, W., Van Berkum, E., 2005. Analyzing highway flow patterns using cluster analysis, *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems*. IEEE, pp. 308-313.

Xia, J., Chen, M., 2007. A nested clustering technique for freeway operating condition classification. *Computer-Aided Civil and Infrastructure Engineering* 22(6), 430-437.

Xia, J., Huang, W., Guo, J., 2012. A clustering approach to online freeway traffic state identification using ITS data. *KSCE Journal of Civil Engineering* 16(3), 426-432.

Zerhari, B., Lahcen, A.A., Mouline, S., 2015. Big data clustering: Algorithms and challenges, *International Conference on Big Data, Cloud and Applications*, Tetuan, Morocco.

Zheng, B., Chen, J., Xia, S., Jin, Y., 2008. Data analysis of vessel traffic flow using clustering algorithms, *Proceedings of the 2008 International Conference on Intelligent Computation Technology and Automation (ICICTA)*. IEEE, pp. 243-246.

Zheng, N., Rérat, G., Geroliminis, N., 2016. Time-dependent area-based pricing for multimodal systems with heterogeneous users in an agent-based environment. *Transp. Res. Part C* 62, 133-148.

Zhong, R., Chen, C., Chow, A.H., Pan, T., Yuan, F., He, Z., 2015. Automatic calibration of fundamental diagram for first‐order macroscopic freeway traffic models. *Journal of Advanced Transportation* 50(3), 363-385.

Appendix

Appendix A. The standard k-means algorithm

Step 1. Randomly choose k initial cluster centers $\{c_1, c_2, \dots, c_k\}$ from n observations $\{x_1, x_2, \dots, x_n\}$.
Step 2. Assign the point x_i $(i = 1, 2, \dots, n)$ to the cluster C_j $(j = 1,2, \dots, k)$ if

$$\|x_i - c_j\| \leq \|x_i - c_l\| \ (l = 1,2, \dots, k, l \neq j) \tag{A.1}$$

Step 3. Update the cluster centers $\{c_1^*, c_2^*, \dots, c_k^*\}$ using

$$c_j^* = \sum (x_i | x_i \in C_j) / n_j \tag{A.2}$$

where n_j is the number of observations in the cluster C_j.
Step 4. If $c_j^* = c_j$ $(j = 1,2, \dots, k)$ or a predetermined maximum number of iterations is reached, terminate the algorithm; otherwise, go to Step 2.

Appendix B. Hierarchical clustering

Step 1. Compute the similarity measure between each pair of objects.
Step 2. The pair of objects in closest proximity is set as a newly formed binary cluster (i.e. a new object in place of the original pair).
Step 3. Repeat Step 1 and 2 until a hierarchical tree is formed.
Step 4. Determine where to cut off the hierarchical tree into clusters (i.e. assign all the objects below each cut to the corresponding cluster).