Transportation Research Part C 82 (2017) 337–357

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Understanding the distribution characteristics of bus speed based on geocoded data

Yuchuan Du [a,*], Fuwen Deng [a], Feixiong Liao [b], Yuxiong Ji [a]

[a] Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China
[b] Urban Planning Group, Eindhoven University of Technology, Eindhoven 513, 5600 MB, The Netherlands

ARTICLE INFO

Article history:
Received 19 September 2016
Received in revised form 8 July 2017
Accepted 11 July 2017
Available online 20 July 2017

Keywords:
Geocoded data
Bus operation
Speed distribution
Finite mixture model
Cluster analysis

ABSTRACT

Data-driven traffic management and control has attracted much attention recently. This paper conducts a series of coherent analyses based on geocoded data to understand the distribution characteristics of bus operational speed and to explore the potential applications of speed distributions. First, an original bipartite model is adopted for capturing instantaneous speed where the suspended and moving states are considered separately and a two-component mixed Weibull distribution is used to model the speed distribution in moving states. The mixed Gaussian distribution with variable components is found to be capable of expressing the speed distribution patterns of different road sections. Second, elaborate analyses on the basis of speed distribution modelling are conducted: (i) regression analyses are conducted to explore the correlations between parameters of instantaneous speed distributions and traffic related factors; (ii) a powerful clustering method using Kullback-Leibler divergence as the distance measure is proposed to grade the road sections of a bus route. These results can be utilized in fields such as bus operations management, bus priority signal control and infrastructure transformation aiming to improve the efficiency of bus operations systems.

1. Introduction

Vehicle speed is one of the most important measurements of traffic performance (May, 1990). Speed distribution is required for many traffic engineering applications. For example, an appropriate speed distribution model is the fundamental input for vehicle generation in traffic microsimulation systems (e.g., Park and Schneeberger, 2003; Llorca et al., 2015) and activity-travel scheduling simulation applications (e.g., Liao et al., 2013; Liao, 2016). Also, speed distribution can be utilized in theoretical analyses of traffic flow characteristics and to devise appropriate traffic operational measures (e.g., Yu and Abdel-Aty, 2014). Specifically, understanding the characteristics of bus speed distribution is essential for public transit operations management. There are two key problems in the investigation of speed distribution: (1) to develop an appropriate model accurately representing the speed distribution; and (2) to conduct analysis on the distribution characteristics, i.e., explaining the distribution patterns or parameters.

To address the first problem, numerous studies have been conducted to model the vehicle speed distributions by applying different statistic distributions, such as normal, log-normal, Student's *t*-distribution and diverse mixture distributions. Leong (1968) and McLean (1978) analyzed free speeds measured on two-lane two-way rural highways and concluded that car

* Corresponding author.
E-mail address: ycdu@tongji.edu.cn (Y. Du).

http://dx.doi.org/10.1016/j.trc.2017.07.004
0968-090X/

speed distributions could be represented by a standardized normal distribution. However, more studies have subsequently shown that normal distribution is only applicable under homogenous traffic flows. For example, Lindner (1965) pointed out that normal distribution could only be applied in limited cases. Instead, a method was introduced in order to represent the non-normal distribution, allowing the approximation of the distribution using an expansion series with successive derivations of the normal distribution. Gerlough and Huber (1975) proposed utilizing log-normal distribution to reflect the skewness of speed distribution. Dey et al. (2006) found that speed distribution on highway could be unimodal or bimodal depending upon the speed variations of different vehicle categories. Consequently, they defined a spread ratio and found out that the speed data followed a unimodal curve only when the spread ratio was in the range of 0.69–1.35. Park et al. (2010) explored the applicability of finite mixtures of normal distributions to capture the heterogeneity in vehicle speed data, and adopted Bayesian estimation via Markov Chain Monte Carlo sampling for parameter estimation. Jun (2010) evaluated the traffic congestion patterns during the Thanksgiving holiday period using a Gaussian mixture speed distribution. Zou and Zhang (2011) investigated the suitability of mixture models using skew-normal and skew-*t* distributions which can fundamentally accommodate skewness and excess kurtosis. Expectation Maximization type algorithm was adopted to estimate distribution parameters. Rossi et al. (2014) compared the performances of different methods, including four-component normal, two-component skew-normal and two-component skew-*t* mixture models. Wang et al. (2012) proposed truncated normal and lognormal distributions to describe vehicle speed distribution because any traffic quantity of interest makes sense only within a limited value range.

With regard to second problem, relatively less research has focused on explaining the practical meanings of vehicle speed distributions or applying these results in the specific contexts. Ko and Guensler (2005) proposed that a mixed speed distribution over a given time period includes congested and uncongested cases. Based on that assumption, they showed it was possible to identify congestion characteristics by exploring the distribution parameters. Park et al. (2010) found that speed separation mainly are resulted from mixed vehicle composition for the flow level, rather than different time periods of a day. Maurya et al. (2015) compared the applicability of several distributions for different classes of vehicles and concluded that the log-normal distribution fitted well with motorized two-wheeled vehicle speed whereas the beta-distribution seemed to be a better fit for speeds of cars, buses and light commercial vehicles. Kyriakopoulou et al. (2016) formulated a methodology to obtain quantitative congestion measures and roadway performance on the basis of the speed distribution parameters.

Most of these aforementioned studies take the mixture speeds of various vehicle types on highways as the research focus, apart from a few cases considering urban roads. However, few studies analyze the operation speed distribution of city buses specifically. In fact, there are significant operational differences between buses and other types of vehicles, e.g., on operational stops and vehicle driving performances, which lead to significant speed distribution differences. Urban ground public transit is a vital infrastructure for cities and the lifeline of urban transportation in modern metropolises. "Transit metropolis" is a development goal for an increasing number of cities, since public transit is conducive to easing traffic congestion and reducing exhaust emissions. Therefore, the enhancement of public transit service, and particularly improving bus speeds in congested areas is an important consideration. The foundation of such work offers sound understanding of bus speed distribution characteristics.

In this study, bus instantaneous and section speed were calculated and analyzed based on GPS-data, which were recorded by GPS devices on buses in the city of Shanghai (China). An original bipartite model was proposed for the instantaneous speed distribution to solve the zero-inflation problem caused by operational stops and congestion delays in the downtown area. In this model, the zero and nonzero components were independently considered. For the nonzero component, a comparison between multiple distribution types was performed, and a two-component Weibull mixture distribution is proposed. A mixed Gaussian distribution model was proposed for the section speed distribution. The Expectation-Maximization algorithm was used for parameter estimation of the proposed models. In addition, simple regression analysis was used to effectively capture the correlation between model parameters and road traffic attributes, considering road section length, green ratio, red light duration in one period, and daily stops. This paper also proposed a powerful and flexible clustering technique using Kullback-Leibler divergence as the measurement of different statistical distributions, which has potential applications in rating bus operating status for different road sections.

The remainder of this paper is organized as follows. Section 2 presents data processing methods and exploratory data analysis results. Section 3 describes the methodology for modelling instantaneous and section speed distributions. Section 4 drills into the analyses of speed distribution parameters and morphology. Section 5 is a further discussion of the research on public transit. Section 6 provides a summary of conclusions and planned future work.

2. Data description and preprocessing

2.1. Data sources

Field data were collected by GPS devices installed on buses in Shanghai from 15th February 2016 to 10th April 2016. The frequency of data recording is 10 s, i.e., a GPS device records positioning information in every 10 s. The data format is shown in the right-hand side of Fig. 1. The study area was limited to 21 sections of Xizang Road located in the central urban area of Shanghai. In this context, a road section is defined as the roadway between two neighboring signalized intersections. The

Fig. 1. Studied road sections and GPS-data format.

included road sections are shown in the left-hand side of Fig. 1. Each intersection is labeled by the name of the crossing road. In total, there are 5,670,795 records after filtering those out of the studied range.

In the study area, there are 40 transit lines travelling through Xizang Road. Showing all the space-time diagrams of the lines results to confusion. Hence, we select one bus line (Line 18) as an example. Fig. 2 shows the space-time diagram of Line 18 during the morning peak (7:00–10:00) and evening peak (16:00–19:00) on 15th February 2016. The blue curves indicate

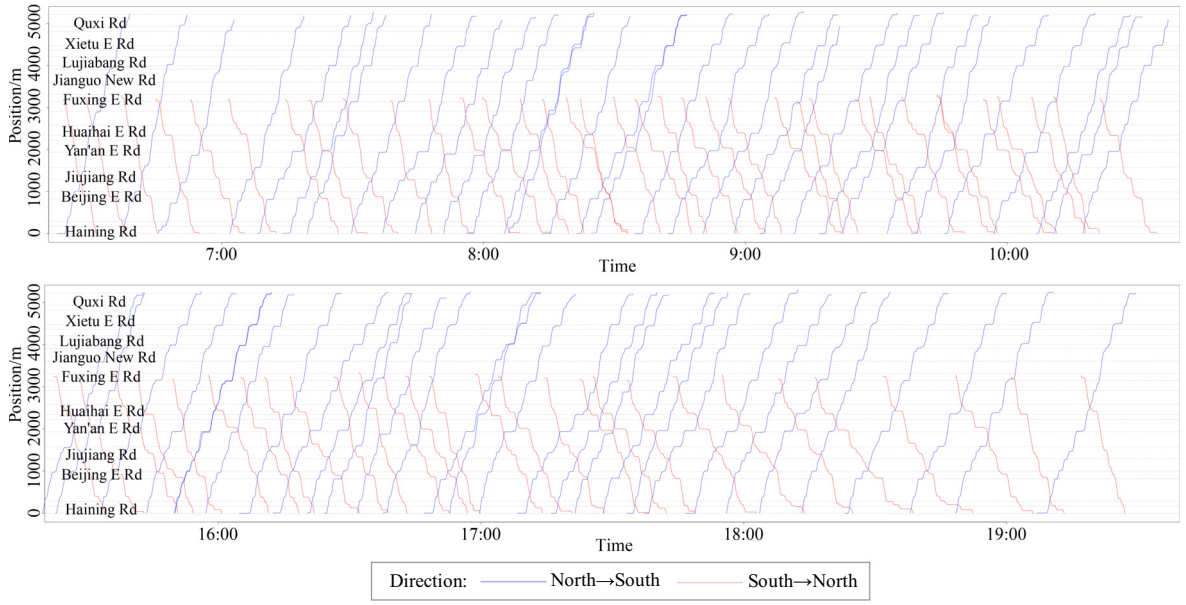


Fig. 2. Space-time diagram of bus Line 18 in Shanghai.

direction from North to South, while the red curves indicate direction from South to North. The bus routes with the direction from South to North don't cover all road sections of the bus corridor according to the transit schedules. As shown in this figure, there are significant differences in speed when the bus travels through different road sections.

2.2. Speed extraction from GPS-data

To obtain the speed data required for the study, two crucial issues need to be addressed: (a) to identify the road sections based on the geocoded points; and (b) to derive the instantaneous speed at each point and calculate the section speed when each bus passes through a road section.

Issue (a) is essentially a map matching problem and has been widely researched (e.g., Brakatsoulas et al., 2005; Quddus et al., 2007). All relevant studies assumed that detailed road network configuration is available. The present paper developed a simple and convenient method to label the geocoded points with road section indices, relying only on the geographical coordinates of intersections. The raw GPS-data provide position and travelling direction information including longitude, latitude and azimuth. For the jth point of an arbitrary bus track, the status information of the bus can be expressed as a 3-tuple: (lon_j, lat_j, azi_j). Suppose that the geographical coordinators of the upstream and downstream intersection of road section i are (lon_k, lat_k) and (lon_{k+1}, lat_{k+1}) respectively. The road section index corresponding to geocoded point j can be determined using the following method (see Fig.3(a)).

Step1: For each road section i, calculate the distances between point j and the upstream intersection (d_k) and the downstream intersection (d_{k+1}) respectively. Measuring distance using longtitude and latitude has a high computational complexity (Ghilani and Wolf, 2007). In this paper, an approximate formula deduced by spherical geometry is applied as Eqs. (1) and (2).

$$d_k = R_e \cdot \sqrt{[\cos(lat_j) \cdot (lon_j - lon_k)]^2 + (lat_j - lat_k)^2} \tag{1}$$

$$d_{k+1} = R_e \cdot \sqrt{[\cos(lat_j) \cdot (lon_j - lon_{k+1})]^2 + (lat_j - lat_{k+1})^2} \tag{2}$$

where R_e represents the radius of the earth.

Step2: For each road section i, calculate the included angle θ_{ij} and perpendicular distance h_{ij} from point j to road section i. Equations are deduced according to the geometrical relation of triangle.

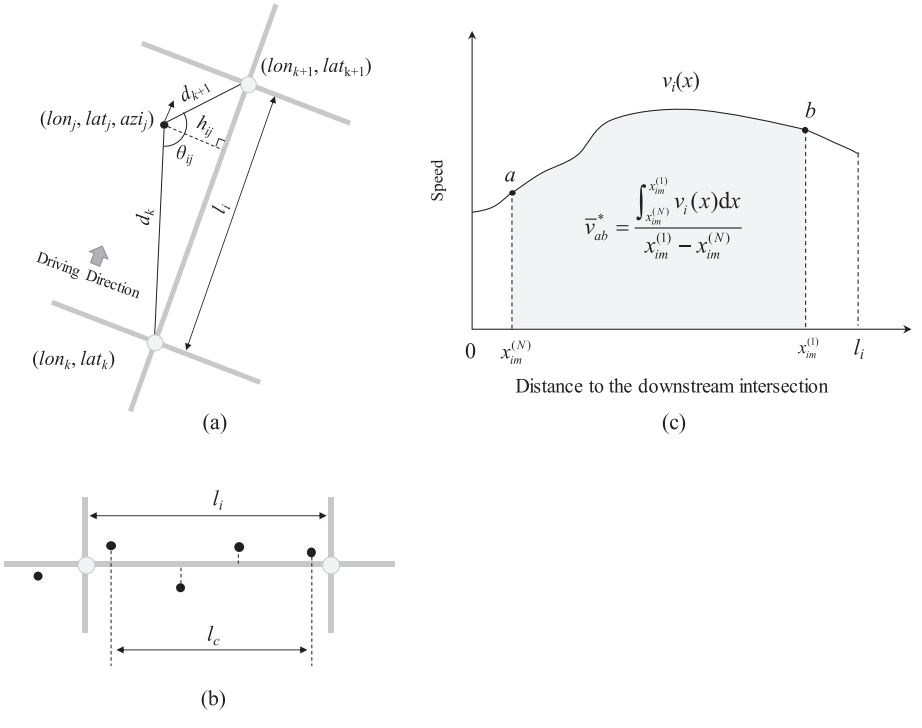


Fig. 3. Strategy of map matching and speed calculation: (a) Map matching method; (b) Effective length for section speed calculation; (c) Spatial distribution function of speed.

$$\theta_{ij} = \frac{d_k^2 + d_{k+1}^2 - l_i^2}{2 \cdot d_k \cdot d_{k+1}} \tag{3}$$

$$h_{ij} = \frac{2 \cdot \sqrt{p_{ij}(p_{ij} - d_k)(p_{ij} - d_{k+1})}}{l_i}, \tag{4}$$

where l_i represents the length of road section i and can be calculated by Eq. (5); p_{ij} can be calculated by Eq. (6).

$$l_i = R_e \cdot \sqrt{[\cos(lat_k) \cdot (lon_k - lon_{k+1})]^2 + (lat_k - lat_{k+1})^2} \tag{5}$$

$$p_{ij} = \frac{d_k + d_{k+1} + l_i}{2} \tag{6}$$

Step3: Determine whether or not the following criteria are satisfied: $\theta_{ij} > 90°$, $h_{ij} < 200$ m (maximum error of the used GPS devices) and $|azi_j - azi_i| < 45°$, where azi_i represents the azimuth of road section i. (The two directions on the road are considered as different road sections.) If so, the geocoded point can be labeled by index i and vice versa.

For Issue (b), the instantaneous speeds were available directly from the raw receiver output. The GPS receivers calculate the instantaneous speed using the Doppler measurement, which is of high accuracy for velocity measurement of non-uniform moving objects (Szarmes et al., 1997; Marchal et al., 2011; Shen and Stopher, 2014). Thus, we only need to derive section speeds. After the map matching process, a single pass track through an arbitrary road section of an arbitrary bus can be extracted. Assume that the distance between the first and last point of the track is l_c, and the total length of the road section is l_i (see Fig. 3(b)). These points were projected onto the road, and the section speed is just l_c divided by t_c (travel time of the segment). Advanced treatments for speed calculations were proposed in previous literature (e.g. Marchal et al., 2011), but no significant differences in the results are observed after a comparison with the present strategy. In order to improve the accuracy, the data fragment was used to calculate the section speed only when $l_c/l_i > r_t$. Obviously, the ratio r_t should be large enough to ensure the accuracy, but an overlarge r_t results in insufficient effective data. This paper set $r_t = 0.85$ as a tradeoff between accuracy and quantity. Fig. 4 depicts the empirical distribution of r_t and the cumulative frequency of speed data. After filtering in accordance with the criterion $l_c/l_i > 0.85$., around 42.8% of the data is kept, as shown in Fig. 4. A larger r_t value leads to a higher accuracy but less amount of speed data.

Also, the calculated section speed must be revised according to the empirical spatial distribution function of the corresponding road segment $v(x)$ (see Fig.3(c)). Suppose the track of the mth bus on the ith road section has N geocoded points. Suppose also that the distance between the downstream intersection and the first point is $x_{im}^{(1)}$, the distance between the downstream intersection and the Nth point is $x_{im}^{(N)}$, with times $t_{im}^{(1)}$ and $t_{im}^{(N)}$ respectively. Then, the adjusted section speed can be expressed as:

$$\bar{v}_{im} = \frac{\bar{v}_i^* \left(x_{im}^{(1)} - x_{im}^{(N)}\right)^2}{\left(t_{im}^{(1)} - t_{im}^{(N)}\right)\left(\int_{x_{im}^{(1)}}^{l_i} v(x)\mathrm{d}x - \int_{x_{im}^{(N)}}^{l_i} v(x)\mathrm{d}x\right)}, \tag{7}$$

where $\bar{v}_i^*$ represents the average speed of the entire road section:

$$\bar{v}_i^* = \frac{\int_0^{l_i} v_i(x)\mathrm{d}x}{l_i}. \tag{8}$$

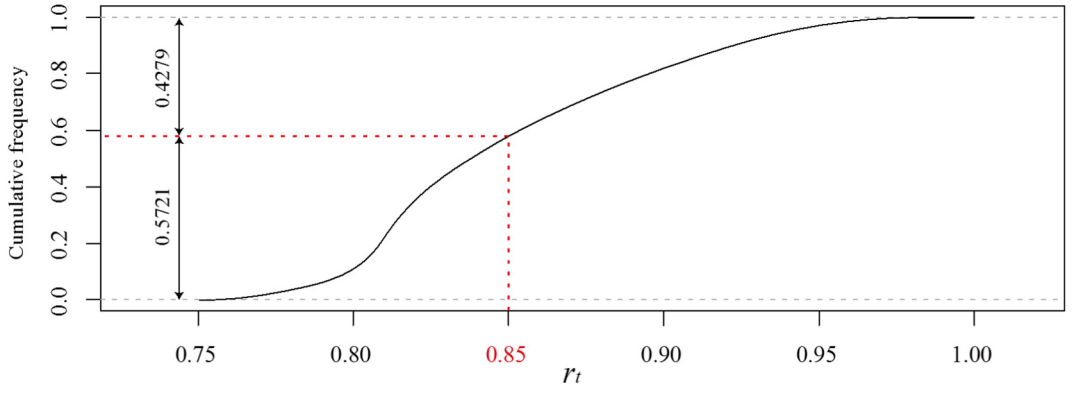


Fig. 4. Empirical distribution function of r_t.

The above process generates two datasets: instantaneous and section speed data for each road section. These datasets are the foundation for the study of bus speed distributions.

2.3. Preliminary analysis

Exploratory analysis is performed to verify the overall properties. First, the temporal distribution of speed is considered. Fig. 5 (a) shows the daily average speed variations of the whole researched road sections, with the time-series decomposition diagram. The time-series decomposition in this study applies an additive model, specified as:

$$Y[t] = T[t] + S[t] + e[t], \tag{9}$$

where T, S and e are the trend, seasonal and error component respectively. The algorithm first determines the trend component using moving average (the period is set to one week) and removes it from the time series. Then, the seasonal component

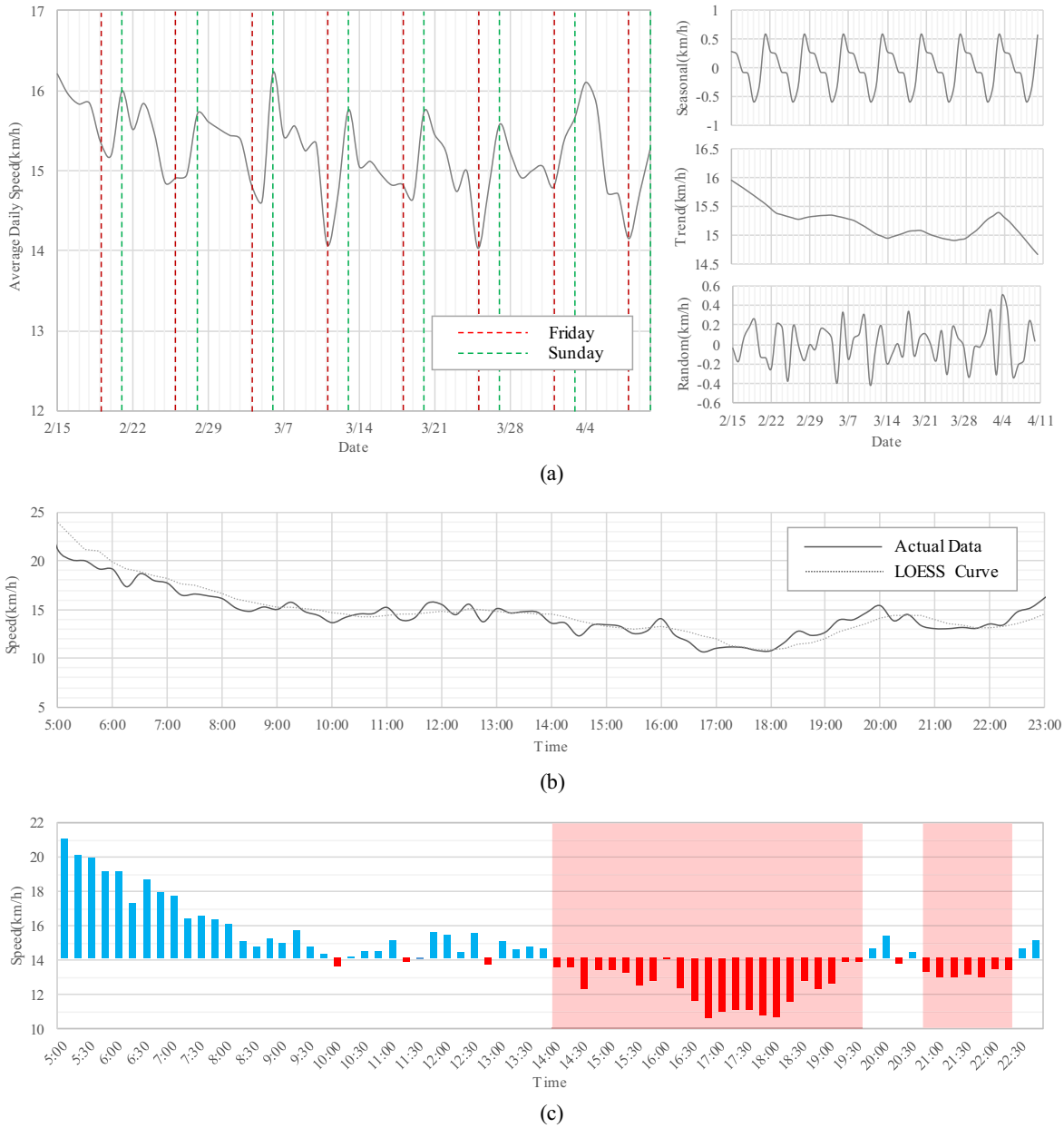


Fig. 5. Temporal distribution of speed data: (a) Daily average speed, (b) Diurnal bus speed, and (c) Quarterly average speed in one day.

is computed by averaging, for each time unit, over all periods. Finally, the error component is determined by removing the trend and seasonal component from the original time series (Kendall et al., 1968). The following conclusions can be summarized from this diagram:

- The daily average speed in most days varies from 14 to 16 km/h.
- The speed data is clearly periodical with a cycle of one week. The peak value appears every Sunday (except April 4) while the minimum value appears every Friday. The difference between the peak and the mininum is approximately 1.2 km/h.
- Speed over the study timeframe shows a slight downtrend, but the trend is not significant. The downtrend is closely connected with the end of the Chinese Spring Festival and the uptrend implies the beginning of the Tomb-sweeping Festival.

Fig. 5(b) shows the diurnal bus speed change for April 8, 2016 (Friday). The intraday lowest speed appears between 16:30 pm and 19:00 pm. The pattern is more striking in Fig. 5(c), showing the difference between the daily and quarterly average speed.

Combining the time and space dimension, the space-time maps of speed are plotted (see Fig. 6). Fig. 6(a) presents the spatial-temporal speed distribution of all the researched road sections and Fig. 6(b) focuses on a single (typical) road section. With the help of space-time maps, slow-moving locations and periods can be readily detected. For instance, the bus-stop in Jiujiang Rd - People Avenue (S→N) road section is a source of significant stagnation, which showed a positive correlation between intersection delay and stop delay.

3. Speed distribution modelling

3.1. Methodology

To describe the speed distribution, this paper adopts the finite mixture modelling method. Compared with the traditional models such as normal or log-normal distributions, finite mixture distributions have greater capability to exhibit multimodality, skewness and excess kurtosis (peakness) (Frühwirth-Schnatter, 2006; Park et al., 2010).

Consider a random variable X that follows a finite mixture distribution. The probability density function (PDF) $p(x)$ can be represented in a mixture density form as follows (Titterington et al., 1985; Frühwirth-Schnatter, 2006):

$$p(x) = \eta_1 p_1(x) + \ldots + \eta_K p_K(x), \tag{10}$$

where K is the number of subgroups, $p_k(x)$ is the PDF of the kth component, and the parameters $\eta_1, \ldots, \eta_K$ are component weights, such that,

$$\eta_K \geqslant 0, \quad \sum_{k=1}^{K} \eta_k = 1. \tag{11}$$

Particularly, assume that all component densities follow the same distribution family $\Gamma(\theta)$ with density $p(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter set of the distribution. Then, the mixture density function $p(x|\boldsymbol{\theta})$ takes the following form:

$$p(x|\theta) = \eta_1 p(x|\theta_1) + \ldots + \eta_K p(x|\theta_K) \tag{12}$$

The Expectation-Maximization (EM) algorithm was adopted for parameter estimation. The EM algorithm is an iterative method for finding maximum likelihood estimators of parameters in statistical models with latent variables (Mclachlan and Krishnan, 2007). The algorithm has become widely used for estimation of univariate and multivariate mixture models due to its conciseness and reliability (Frühwirth-Schnatter, 2006). The iterative process of EM algorithm can be expressed as follows.

Loop iteration (Until convergence conditions are satisfied) {

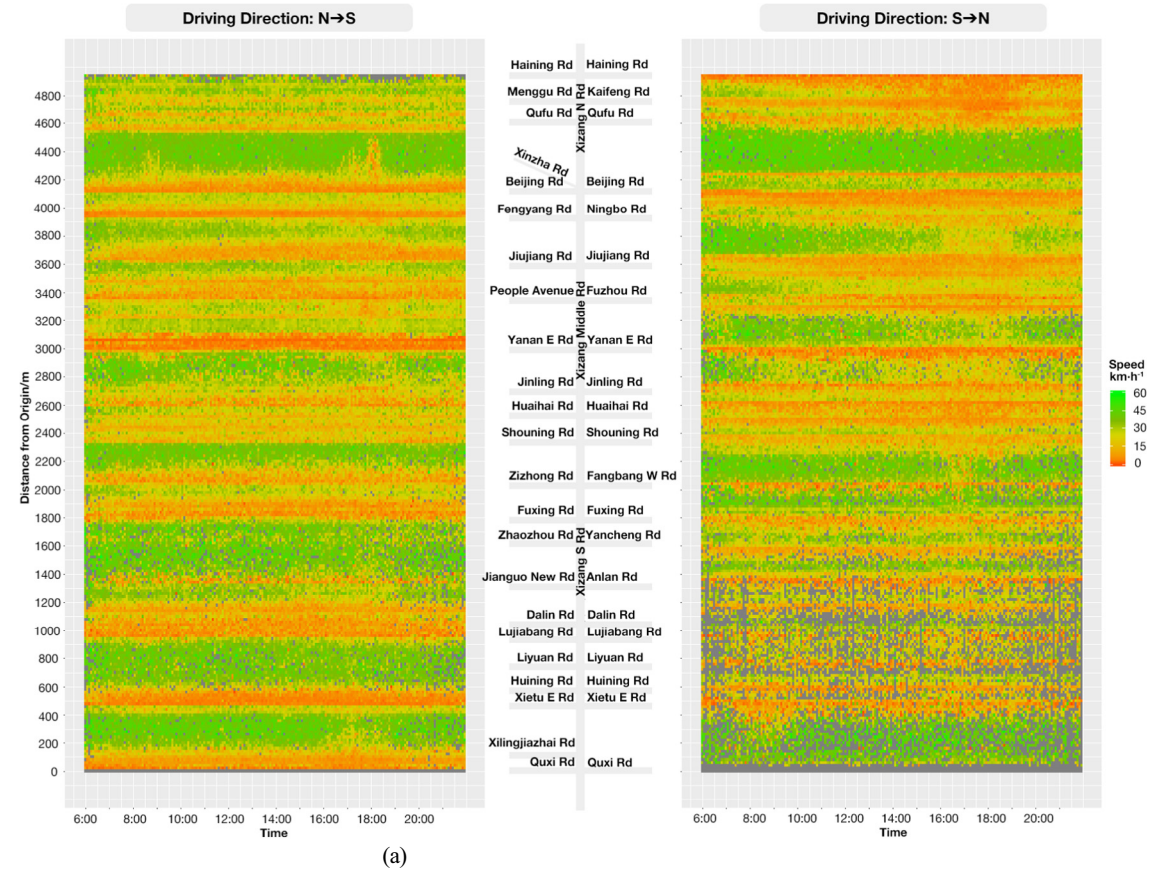
(E-Step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta) \tag{13}$$

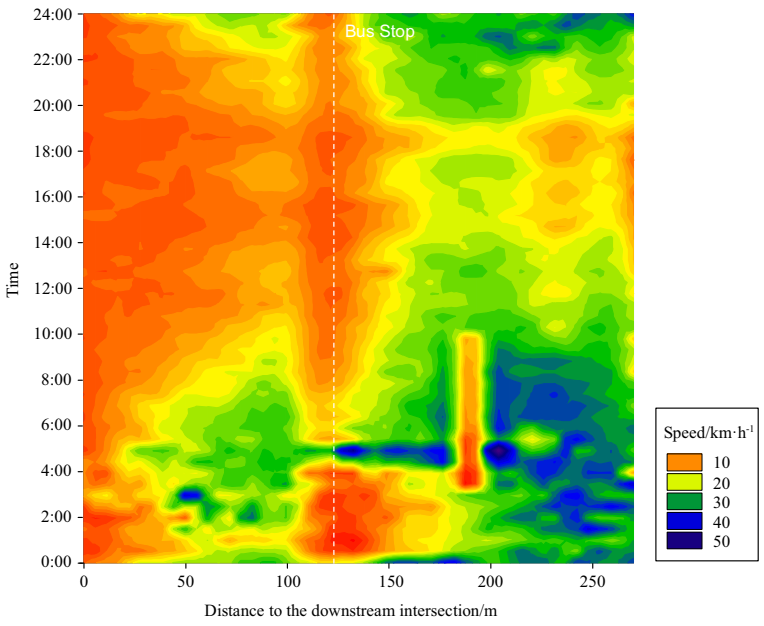
(M-Step) Set

$$\theta := \arg\max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \tag{14}$$

}



(a)



(b)

Fig. 6. Spatial-temporal distribution of speed data: (a) All included road sections and (b) Jiujiang Rd-People Avenue (S → N) road section.

In Eqs. (13) and (14), z is the latent variable and $Q(z)$ is the prior distribution of z. In addition, two extensions of the EM algorithm, i.e. the Expectation/Conditional Maximization algorithm and the Expectation/Conditional Maximization Either algorithm, can deliver improved efficiency (Zou and Zhang, 2011).

3.2. Instantaneous speed distribution

In contrast to ordinary cars, buses tend toward a lower average speed because of the frequent operational stops and worse vehicle performance (a general situation in China), which is particularly true in the city centers. Therefore, the bus instantaneous speed distribution is significantly different from that of car, which is reflected in a high zero probability density, i.e. the probability of zero speed. This fact causes difficulties for distribution fitting. To address this problem, a bipartite model is introduced where the zero and the nonzero components are independently considered. Due to inherent drift of GPS positioning, even if a bus actually stops, a small value of speed is still recored. Thus, the model sets a small threshold; and speeds smaller than, greater than or equal to the threshold are handled separately.

Therefore, the following PDF was adopted to describe the instantaneous speed distribution of buses,

$$p(x) = \begin{cases} p_0 & x < \delta \\ (1-p_0)g(x-\delta;\theta) & x \geqslant \delta \end{cases}, \tag{15}$$

where δ is the speed threshold, p_0 is the probability that the speed is smaller than δ, and $g(x)$ is a mixed distribution where the subgroups follow the same distribution family. To select an appropriate distribution family to describe the distribution of each subgroup, the paper compared PDFs of the Gamma, Weibull and log-normal distributions, which are,

$$Gamma(x|\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\beta}}, x > 0 \tag{16}$$

$$Weibull(x|k,\lambda) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k} & x \geqslant 0 \\ 0 & x < 0 \end{cases} \tag{17}$$

$$LNorm(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right) \tag{18}$$

where $\Gamma(x)$ (Eq. (16)) is the Gamma function,

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt. \tag{19}$$

A test sample with 50,000 records was generated by random sampling from the instantaneous speed data of all road sections. Above all, assign δ = 1.5 km/h and calculated p_0 = 0.3301. For the remaining data, parameter estimation was performed using the above distributions. The Kolmogorov-Smirnov (K-S) test value and Akaike information criterion (AIC) were calculated to compare the goodness of fits. Table 1 shows the fitting results of the three mixture models. It turned out that the two-component mixed Weibull distribution has the best fit result, i.e., the minimum K-S test value and AIC, followed by the mixed Gamma distribution and mixed log-normal distributions. Thus, the final outcome was

$$p(x) = \begin{cases} 0.3301 & x < 1.5 \\ 0.6699g(x-1.5) & x \geqslant 1.5 \end{cases}, \tag{20}$$

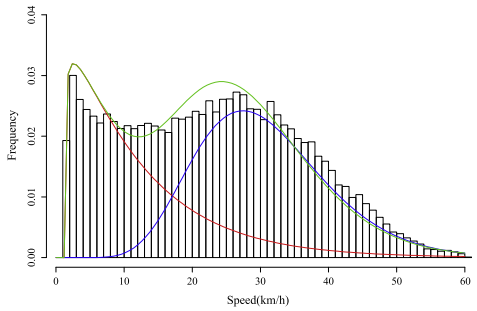
$$g(x) = 0.3272 \cdot \left[0.1119\left(\frac{x}{9.1640}\right)^{0.0253}e^{-\left(\frac{x}{9.1640}\right)^{1.0253}}\right] + 0.6728 \cdot \left[0.0914\left(\frac{x}{30.8033}\right)^{1.8148}e^{-\left(\frac{x}{30.8033}\right)x\geqslant 1.5^{2.8148}}\right]. \tag{21}$$

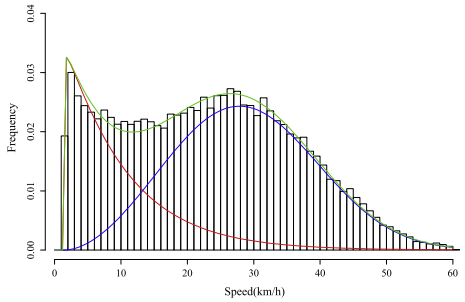
The two-component mixed Weibull distribution was also applied to 21 road sections separately, with the model parameters of the road sections shown in Table A1 in the Appendix. The road sections are numbered according to the order from north to south and only the weight of the first subgroup η_1 is listed as η, since $\Sigma\eta_i = 1$. There are significant differences among instantaneous speed distributions of different road sections, as shown in Fig. 7. We hypothesize that the differences arise from the differences of operating conditions, such as transport facilities, signal controls and operation stops, and then conduct regression analyses to explore the correlations between the distribution parameters and road traffic attributes in Section 4.1.

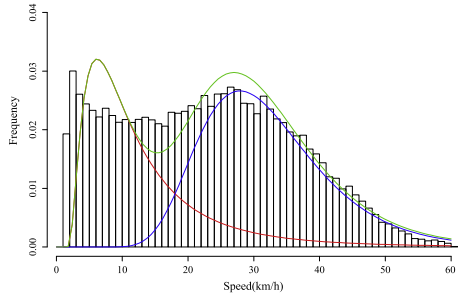
3.3. Section speed distribution

The mixed normal (Gaussian) distribution model was adopted to fit the section speed distribution for each road section. For the instantaneous speed distribution, two components were sufficient due to the uniformity of distribution patterns in every road section. However, the section speed distributions show intense heterogeneity, and a consistent component number may not be applied.

Table 1
Fitting comparison between three mixture models.

Two-component mixed Gamma distribution				
Schema	Parameters		K-S value	AIC
	α_1	1.1031	0.0454	265326.8
	β_1	0.0980		
	η_1	0.4346		
	α_2	8.9030		
	β_2	0.3047		
	η_2	0.5654		

Two-component mixed Weibull distribution				
Schema	Parameters		K-S value	AIC
	k_1	1.0253	0.0379	264703.0
	λ_1	9.1640		
	η_1	0.3272		
	k_2	2.8148		
	λ_2	30.8033		
	η_2	0.6728		

Two-component mixed log-normal distribution				
Schema	Parameters		K-S value	AIC
	μ_1	2.1745	0.0524	266973.6
	σ_1	0.7970		
	η_1	0.4103		
	μ_2	3.3742		
	σ_2	0.3184		
	η_2	0.5897		

Previous studies have tried to finding an appropriate unique number of components to describe the distribution covering the entire dataset. In contrast, this paper uses different number of components for different road sections, with the number determined by a selection criterion.

To select a model for a particular road section, there are two key evaluation criteria: goodness of fit and avoiding over-fitting. In order to determine the component number with the best fitting efficiency, a component selection algorithm based on the criterion of minimizing the estimation error was adopted. The algorithm calculated the Bayesian Information Criterion (BIC) of candidate models with 1 to N_c Gaussian components. The BIC penalizes free parameters more strongly than AIC, thus it tends to produce a smaller number of parameters, and in this context fewer distribution components. However, the algorithm may lead to many arbitrary components if N_c is too large, which was not expected. For this reason, a 5-component bound was set, $N_c = 5$, to reduce computational complexity and avoid over-fitting. Each component in the speed distribution points to a typical working condition of the buses under the influence of multi-factors. The pattern of the whole distribution

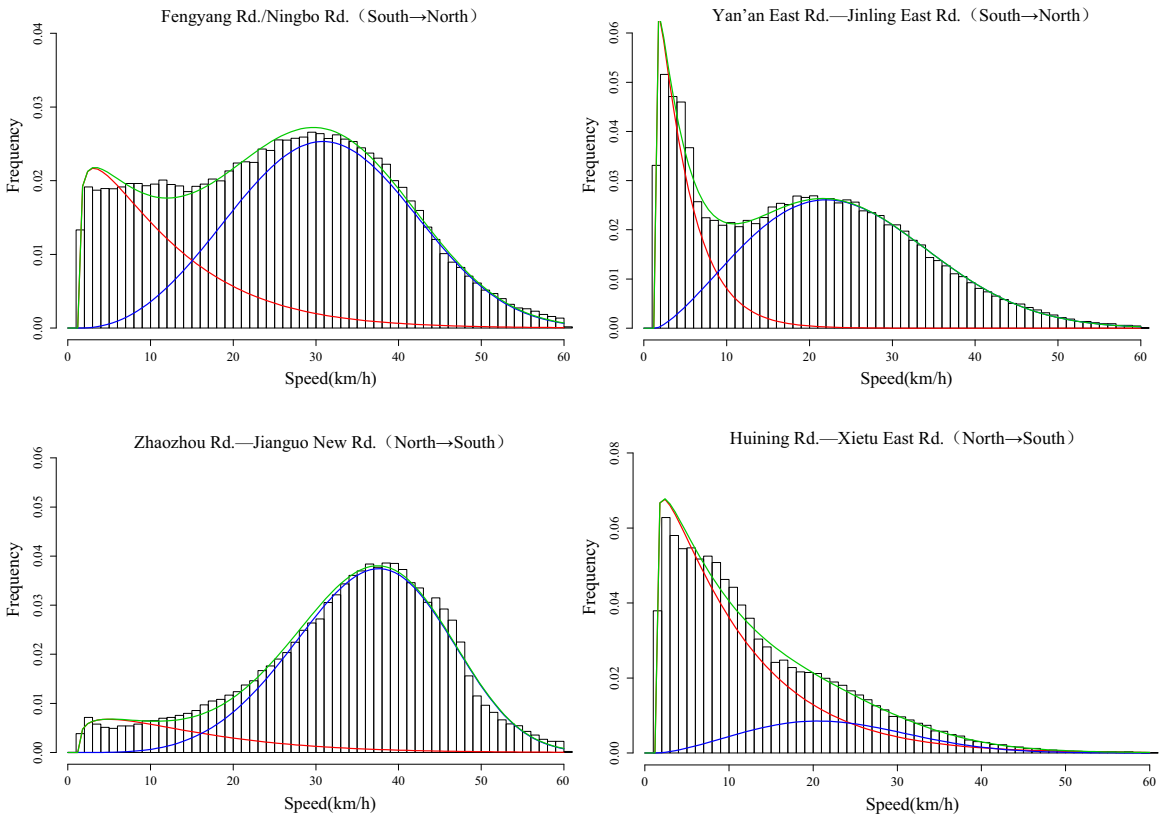


Fig. 7. Instantaneous speed distributions of four road sections (as examples).

for a specific road section represents the speed characteristics including several typical working conditions. As a result, the criterion balances between computation complexity and power of the model well.

Fig. 8 shows the process for the Jiujiang Rd.-People Avenue (Travelling direction: S → N) road section. The first step is parameter estimation for a mixture normal (Gaussian) distribution with 1–5 component(s):

$$p(x) = \sum_{i=1}^{K} \eta_K Norm(x|\mu_K, \sigma_K) \quad K \in \{1,2,3,4,5\}, \tag{22}$$

where K is the number of components, and $Norm(x)$ is the PDF of normal distribution.

Moreover, the BIC is calculated for each model and the model with the minimum BIC value is selected to describe the section speed distribution. The derived model parameters of all the road sections are listed in Table A2 in the Appendix.

A noteworthy finding is that about 74% of the road sections need four or five subcomponents to describe the distribution characteristics, which indicates extreme complexity and variability of speed distributions. We infer that it is due to both the complex traffic flow characteristics in urban roads and the particularity of buses compared with general vehicles. Moreover, there is a great difference between the speed distributions in different road sections, which has implications for finding bottlenecks and carrying out suitable operations management.

4. Elaborate analysis of speed distributions

In this section, further analyses on distribution characteristics were performed. For the instantaneous speed distribution, regression analysis was conducted to explore the correlation between distribution parameters and road traffic attributes. For the section speed distribution, cluster analysis was performed to grade bus operation level on different road sections.

4.1. Distribution parameter factors

After modelling the speed distribution, it is useful to explore the correlations between the model parameters and the operating conditions to understand the influence mechanism of public transit speed. To achieve this, this paper conducted single factor analysis by using the simple linear regression method.

n_c	μ_1	μ_2	μ_3	μ_4	μ_5	σ_1	σ_2	σ_3	σ_4	σ_5	η_1	η_2	η_3	η_4	η_5	BIC
1	6.23	—	—	—	—	2.16	—	—	—	—	1.00	—	—	—	—	69592.91
2	5.53	10.19	—	—	—	1.41	1.20	—	—	—	0.85	0.15	—	—	—	66667.42
3	4.88	7.08	10.04	—	—	1.03	0.52	1.25	—	—	0.61	0.22	0.17	—	—	65661.48
4	4.87	7.10	14.97	10.13	—	1.02	0.55	0.92	0.98	—	0.60	0.24	0.00	0.15	—	65379.17
5	3.98	4.12	5.24	6.95	10.05	0.16	0.81	0.54	0.58	1.24	0.02	0.25	0.27	0.29	0.17	65636.16

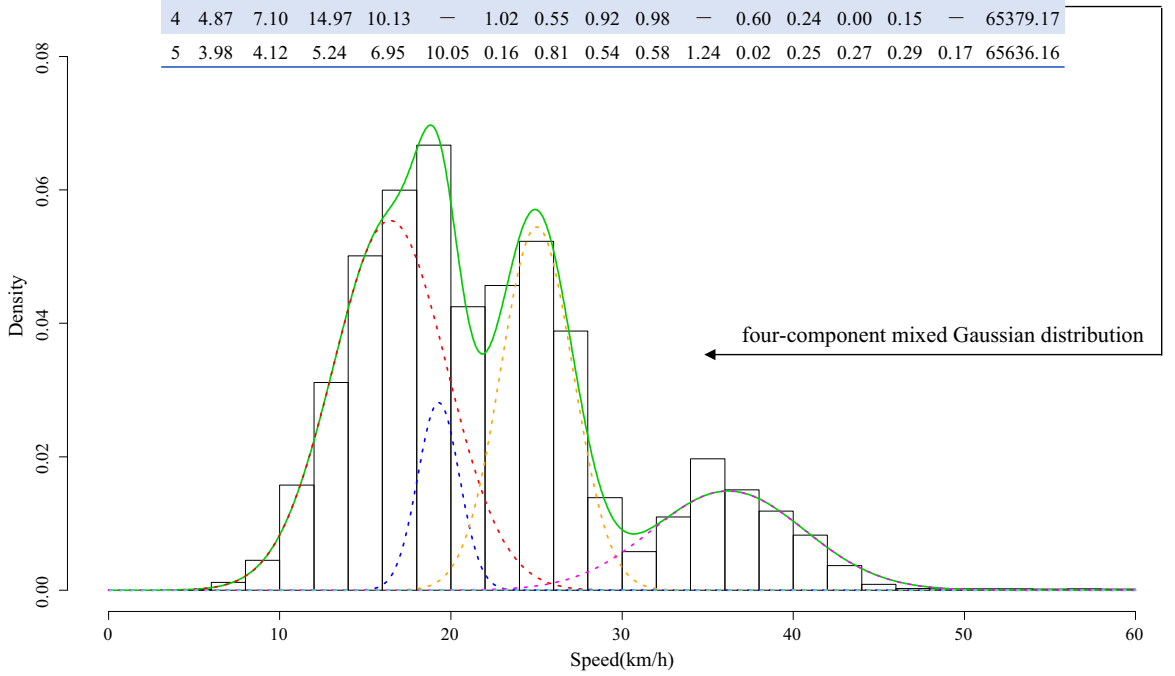


Fig. 8. Example of section speed distribution (Jiujiang Rd. – People Avenue, S → N).

The bus operational speed is influenced by many factors. Besides the actual traffic conditions, the most significant factors are intersection delay and stop delay. This argument is also supported by the literature (Cortés et al., 2011; Ma and Wang, 2014; Zhang et al., 2014). For this reason, road length (l), green ratio (rg), maximum red light duration in one single period (tr) and daily operation stops (s) were chosen as independent variables. The first three variables determine the proportion of the total delay and the last variable is related to the stop delay. The choice of the variables is heavily for signal areas because these are useful for bus signal priority control which contributes to enhancing the bus operation speed. The correlation coefficients between any two chosen variables are less than 0.1. The single factor analysis is mainly from the perspective of convenience for engineering applications.

Note that the instantaneous speed distribution model is a bipartite model. The model parameters include p_0, the weight of each Weibull component η_i and the distribution parameters of each component k_i and λ_i. The mean and variance of Weibull distribution are expressed below,

$$E(X) = \lambda\Gamma\left(1+\frac{1}{k}\right), \tag{23}$$

$$Var(X) = \lambda^2\left[\Gamma\left(1+\frac{2}{k}\right) - \Gamma\left(1+\frac{1}{k}\right)^2\right]. \tag{24}$$

For $k > 1$, the mean value is controlled by λ, while variance is influenced by k and λ. Also, p_0 is an important parameter that reflects the proportion of stagnation and movement. Thus, k, λ and p_0 should all be considered as corresponding variables in the regression model.

Fig. 9 shows the regression analysis results, including the regression lines at 95% confidence intervals. Correlations λ_1 - ~ daily operation stops, $\lambda_2 \sim$ road length, $k_2^* \sim$ green ratio and $p_0 \sim$ green ratio were considered. Fig. 9(a) shows that λ_1 is correlated with daily operational stops in a road section with bus stops (only road sections that have bus stops are plotted in this diagram). Fig. 9(b) reveals that λ_1 is correlated with road length. In Fig. 9(c) and (d), defining $k_2^* = k_2(1-\eta)$, and the correlation between k_2^* and green ratio was explored. Fig. 9(d) incluses only road sections without bus stops, and a strong correlation is evident. Fig. 9(e) and (f) express the negative correlation between p_0 and green ratio (Fig. 9(f) only covers the non-stop road sections).

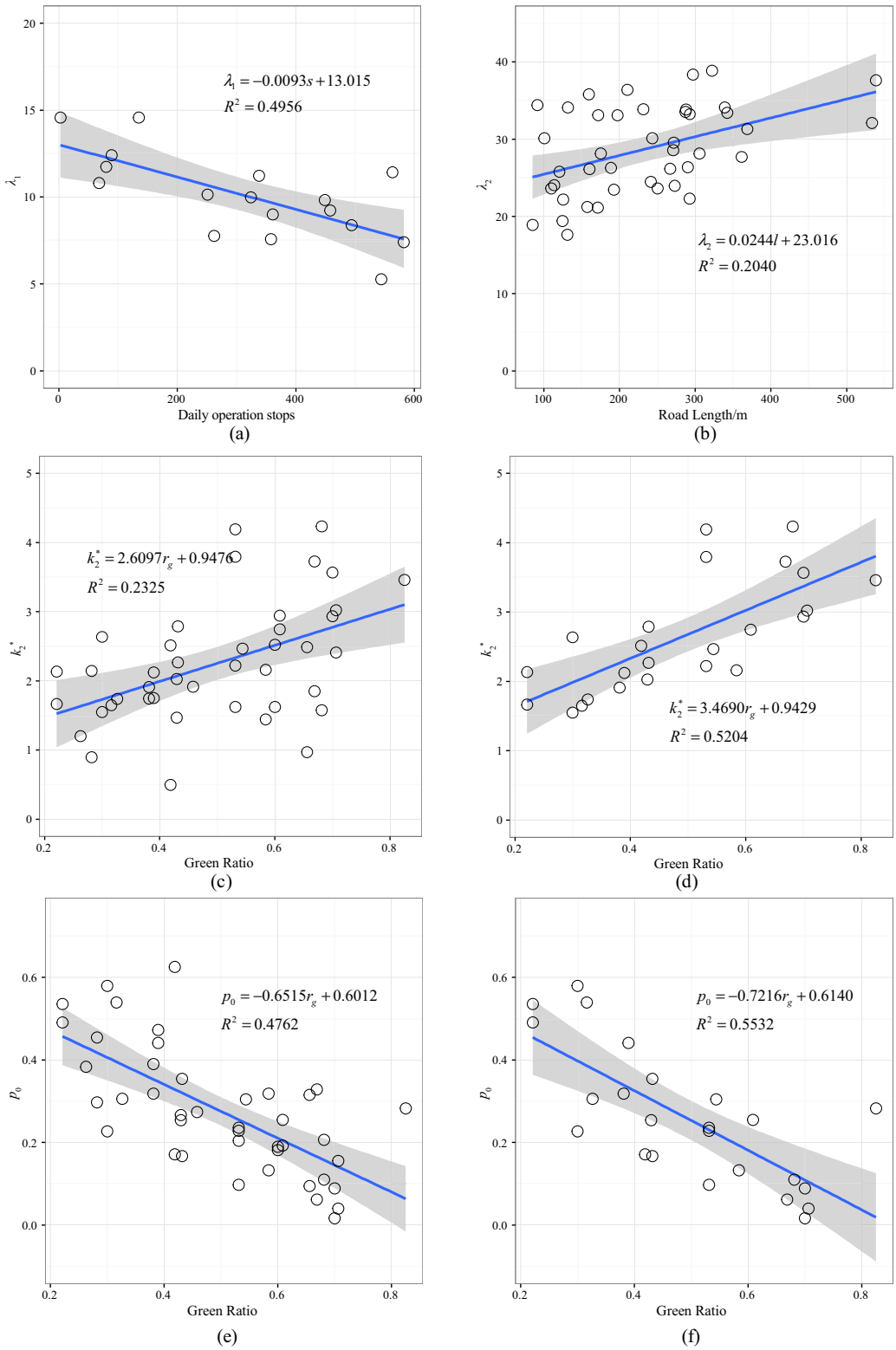


Fig. 9. Regression correlation between model parameters and traffic attributes: (a) $\lambda_1 \sim$ daily operation stops; (b) $\lambda_2 \sim$ road length; (c) $k_2^* \sim$ green ratio; (d) $k_2^* \sim$ green ratio (non-stop road sections); (e) $p_0 \sim$ green ratio; (f) $p_0 \sim$ green ratio (non-stop road sections). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

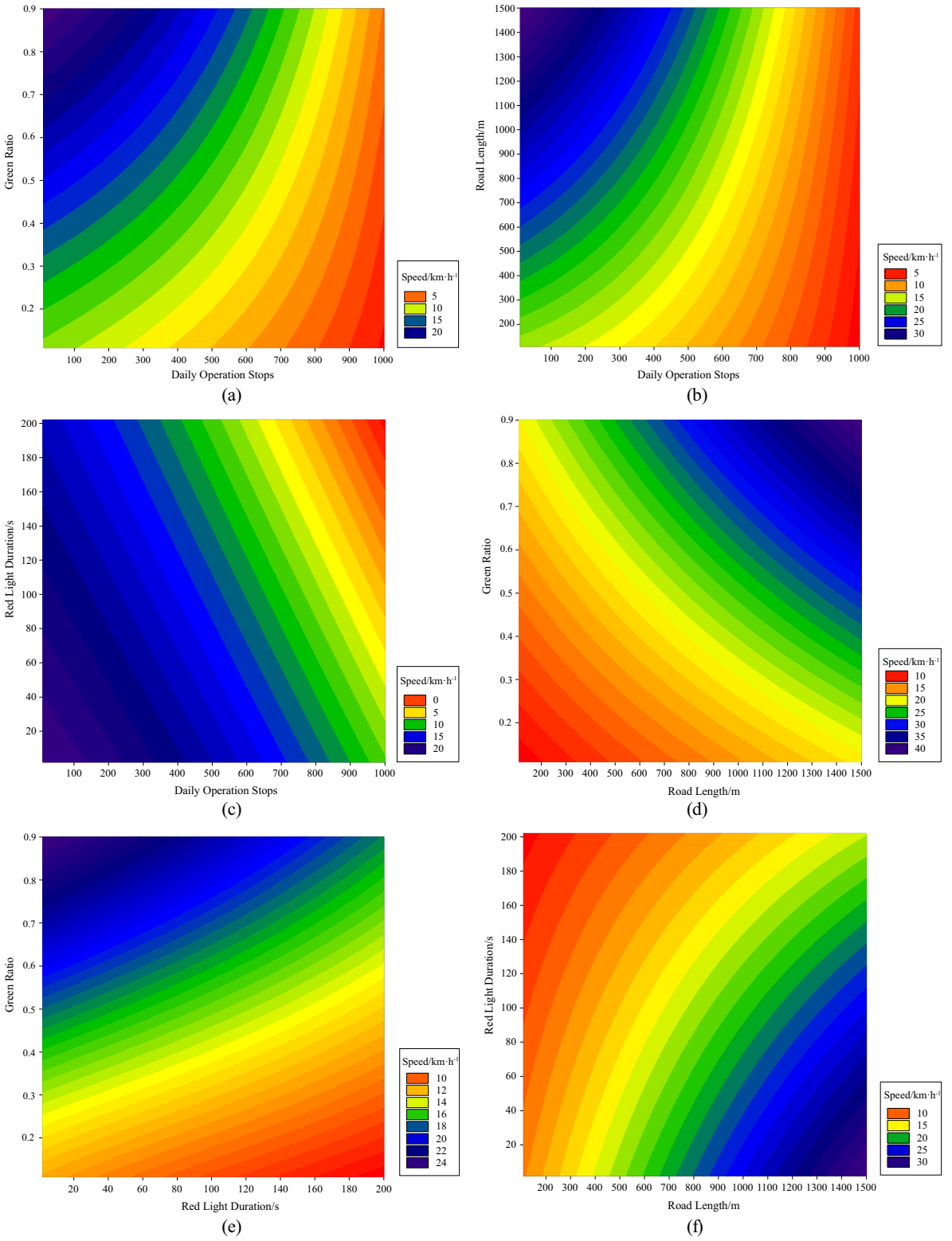


Fig. 10. Speed thermodynamic diagram under a certain boundary condition: (a) $l = 350$ m, $t_r = 90$ s; (b) $r_g = 0.6$, $t_r = 90$ s; (c) $l = 350$ m, $r_g = 0.6$; (d) $t_r = 90$ s, $s = 200$; (e) $l = 350$ m, $s = 200$; (f) $r_g = 0.6$, $s = 200$.

The speed thermodynamic diagrams (see Fig. 10) are plotted on the basis of the above-mentioned correlations. Each subgraph in Fig. 10 shows the trend of speed change with two independent variables under the indicated boundary conditions. This illustrates the magnitude of the various effects for each factor, and is of some significance to transit operation management.

4.2. A rating method based on cluster analysis of speed distributions

Evaluation and ranking of road sections in a bus route is a challenging task for transit operations. Traditionally, statistics of speed data, such as mean, variance, kurtosis and skewness were utilized to rank the bus travel conditions. The method holds under homogenous traffic streams. When the speed distribution becomes multimodal, nevertheless, these criteria only reflect a small part of the travel characteristics and do not have the capability to deal with the ranking problem. Therefore, this paper adopted an original clustering technique using Jensen–Shannon divergence, a modified form of Kullback–Leibler divergence, as the distance measure to address this problem. Given the speed distribution models of each road section, the proposed technique can cluster road sections with similar distribution characteristics.

There are a variety of measurements that can be used as the distances between samples such as Euclidean distance, Mahalanobis distance, cosine similarity and Kullback–Leibler divergence (Kaufman and Rousseeuw, 1990). For the problem in this paper, each sample is a probability distribution or a set of speed data obeying a specific probability distribution. Euclidean distance and Mahalanobis distance cannot measure the distance between probability distributions. If an alternative data set generated by the distributions is used, the "demension disaster" arises. Cosine similarity does not work well because it is closely related to the normal distribution but the data on which it is applied is not close to normal distribution (Whissell and Clarke, 2013). In fact, the collected data is more from a mixture distribution. For this reason, the Kullback–Leibler divergence is used to deal with the problem.

Kullback–Leibler divergence (K-L divergence) is a measure of the difference between two probability distributions (Kullback and Leibler, 1951). For two distributions P and Q of a continuous random variable, K-L divergence is defined as:

$$D(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x. \tag{25}$$

K-L divergence is not symmetric, i.e. $D(P||Q) \neq D(Q||P)$ typically. Therefore, the symmetrized Jensen–Shannon divergence (J-S divergence) (Endres and Schindelin, 2003) was adopted to satisfy the basic nature of distance:

$$JSD(P||Q) = \frac{1}{2} D(P||M) + \frac{1}{2} D(Q||M), \tag{26}$$

where $M = \frac{1}{2}(P + Q)$.

For two Gaussian distributions $\hat{f}$ and $\hat{g}$, K-L divergence has a closed form expression:

$$D(\hat{f}||\hat{g}) = \frac{1}{2}\left[\log \frac{|\Sigma_{\hat{g}}|}{|\Sigma_{\hat{f}}|} + \mathrm{Tr}(\Sigma_{\hat{g}}^{-1}\Sigma_{\hat{f}}) - d + (\mu_{\hat{f}} - \mu_{\hat{g}})^T \Sigma_{\hat{g}}^{-1}(\mu_{\hat{f}} - \mu_{\hat{g}})\right]. \tag{27}$$

However, for mixed distributions, it is typically unable to deduce a closed formed expression. Therefore, the Monte Carlo sampling method (Hershey and Olsen, 2007) was usually employed to calculate K-L divergence. Given two distributions f and g, n independent identically distributed samples $\{x_i\}_{i=1}^n$ were drawn from f, and we obtain the following,

$$D_{\mathrm{MC}}(f||g) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(x_i)}{g(x_i)} \overset{n\to\infty}{\to} D(f||g) \tag{28}$$

D_{MC} is supposed to get close to the true value D by increasing n. In general, the algorithm provides good precision level when $n \geq 10^3$.

The analysis process is detailed below.

Step1: Obtain the PDFs of speed distributions for the researched road sections.
Step2: Calculate the J-S divergence between any two speed distributions δ_{ij} using the Monte Carlo sampling and obtain the distance matrix:

Table 2
Clustering result with statistics.

Cluster	Size	Maximum distance	Average distance	Diameter	Separation
A	5	0.2904	0.1548	0.3749	0.1446
B	18	0.6941	0.2489	0.8016	0.2359
C	15	0.3776	0.2556	0.3343	0.1567
D	4	0.2397	0.1618	0.4647	0.2029

Fig. 11. Clustering result based on section speed distribution.

$$D = \begin{bmatrix} \delta_{11} & \cdots & \delta_{1n} \\ \vdots & \ddots & \vdots \\ \delta_{n1} & \cdots & \delta_{nn} \end{bmatrix}, \tag{29}$$

which satisfies $\delta_{ij} = \delta_{ji}$, $\delta_{ij} = 0$ $(i = j)$.

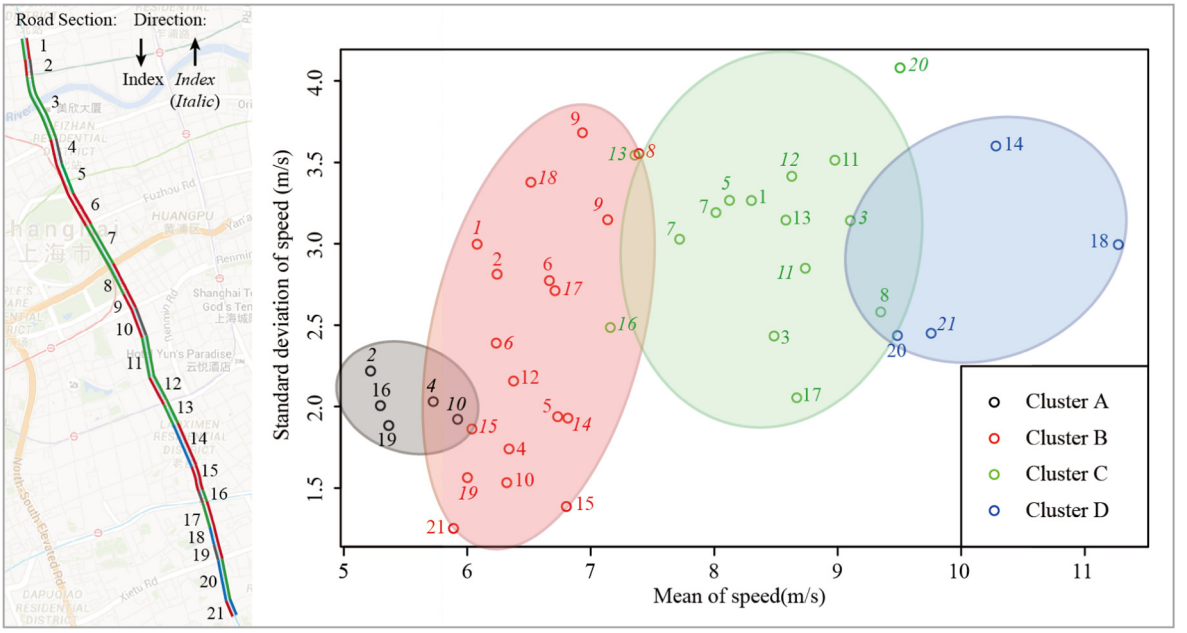


Fig. 12. A specific-dimension visualization of clusters.

Step3: Input the distance matrix into an appropriate clustering method such as K-Medoids (Chen et al., 2010; Park and Jun, 2009). The category quantity can be determined using silhouette coefficients or just according to the research needs.

In this paper, the road section speed distribution was used for cluster analysis, and the basic input was the PDFs of mixed Gaussian distributions. The clustering method works based on the morphological characteristics, therefore it is not affected by discrepancies in the number of Gaussian components. Table 2 and Fig. 11 shows the resultant clusters. In Fig. 12, the clusters are visualized in two specific dimensions (mean and standard deviation). Note that each point in this figure represents a probability distribution.

By clustering, the researched road sections fall into four classes: Levels A, B, C and D, which reflect the different characteristics of the speed distributions. The results can also be treated as the ranking of road sections for bus travel. The features of each class were interpreted as follows:

Level A: Buses drive on these road sections at a very low speed. (Average speed: 19.8 km/h)
Level B: Speed distributions in this category exhibit more heterogeneity than Level A and are typically bimodal or multimodal. Nevertheless, the main peak usually lies in the low speed range. (Average speed: 23.4 km/h)
Level C: Speed distributions reflect the strongest dispersion. The distribution peaks are relatively smooth. (Average speed: 30.4 km/h)
Level D: This category reflects an ideal travelling status. Speed data is congregated in the high speed range. (Average speed: 36.7 km/h)

5. Discussion

The study of bus speed distribution characteristics has been discussed in detail through a case study. As a generalization, the methodology is presented in the following flowchart (Fig. 13), which gives the procedures for analyzing bus speed distribution characteristics.

This study presents a series of analyses for understanding bus speed distribution patterns based on geocoded data collected by on-board GPS receivers in a bus corridor. Although the proposed analytical approach was implemented on a bus corridor, it is applicable to a bus network by considering the bus routes as long bus corridors. Given the geocoded data of the bus trajectories, Fig. 13 shows the complete route map for the research and application of bus spped distributions. First, the raw data should be preprocessed to obtain the desired speed information. A verification process should be carried out at the same time to ensure the validity of the data. Second, both the instaneous and section speed distributions are studied through model selection process and parameter estimation. Third, elaborate analyses of speed distributions are designed to uncover the relationship between the distribution patterns and the actual travelling environment. The outcomes will provide the necessary references for policy development and engineering applications.

Speed distribution is a classical problem in traffic flow theory. Very few studies analyse the operational speed distribution of city buses specifically. The lack of study on the bus speed distributions may cause inconvenience for theoretical research

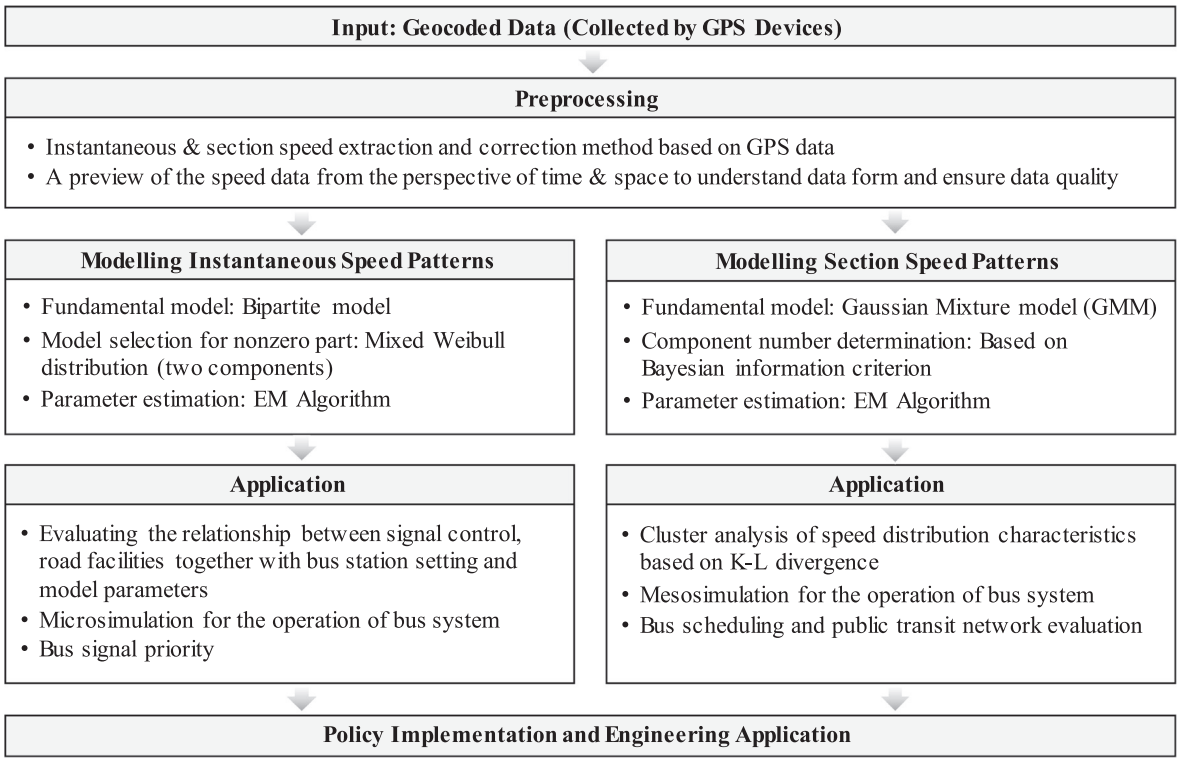


Fig. 13. The procedures for analyzing bus speed characteristics.

and engineering practice as to the public transportation system. For this reason, the specialized study on the characteristics of speed distribution of buses is carried out. Furthermore, with the widespread promotion of public transit, especially in developing countries such as China and India, the research on the operation of bus system is becoming increasingly important.

In the paper, both instaneous speed and section speed were studied. Instaneous speed distribution patterns describe the bus travel characteristics on a micro level and can be used in scenarios which require fine-grained speed characteristics such as microsimulation for an urban bus system, while section speed distribution is applicable for the meso or macro–simulations in contrast. Instantaneous speed distributions in different road sections have the same form, namely two-component mixed Weibull distribution for non-zero data. The physical significance could be normal and slow-speed conditions respectively, to some extent similar to the two-fluid model in transportation systems. However, section speed distributions show significant heterogeneity in different road sections. For a strong descriptive power, GMM (Gaussian Mixture Model) was adopted based on the fact that it is able to fit arbitrary distributions on the basis of Wiener's Tauberian theorem (Rudin, 1973). The key problem here is the determination of the number of sub-components. A component selection algorithm based on the criterion of minimizing the estimation error was adopted with a setting of 5-component bound in order to reduce computational complexity and avoid over-fitting. The study uncovered the absence of a unified model of bus speed distribution. Therfore, in the operations of a public transport system, varied bus management and control strategies, such as specific signal control and holding control, should be implemented in different road sections or areas (Xuan et al., 2011). Bus lanes are recommended to reduce this kind of heterogeneity and improve the stability of bus operations (Eichler and Daganzo, 2006; Yao et al., 2015).

A common application of the speed distribution model is for the vehicle flow generation in traffic simulation systems. Besides, the research is useful for applications which aim to solve different levels of management, control and scheduling problems. For instance, the instantaneous speed distribution model is an input for bus signal priority control, which requires fine-grained speed characteristics; and the section speed distribution model provides a tool for evaluating the bus scheduling and public transit network services. Another potential application of the speed distribution patterns is cognizing the traffic condition and identifying the traffic anomalies in a specific road section (Thajchayapong et al., 2013).

6. Summary and conclusions

Data play an increasingly important role in traffic research. With the popularization of diverse sensors and GPS devices, traffic data collection is becoming increasingly convenient. Traffic related awareness based on big data analysis has also

attracted more attention in recent years. This paper provides practical interpretations of bus speed distributions based on automatic vehicle location data and conducted numerical investigations. Speed distribution models were established to model the patterns of instantaneous and section bus speeds. For the instantaneous speed distribution, a bipartite model was adopted, in which zero and non-zero components were considered separately. The non-zero component was described using the two-component mixed Weibull distribution. For section speed, mixed Gaussian distributions with variable number of components were used to express distribution characteristics of different road sections. These models provide a good descriptive method to express bus speed distributions. Thereafter, speed distribution analysis was performed to explore potential applications. Regression analysis based on the instantaneous speed distribution provided correlations between distribution parameters and several contributing factors, including road length, green ratio, maximum red light duration in a single period and daily operational stops. Meanwhile, a powerful method using Kullback-Leibler divergence as the clustering measure was proposed to rank the road sections on a bus route. These techniques and findings is beneficial to data-driven public transit operations management, as well as transformation of traffic infrastructure aiming to improve transit speed in congested areas.

Nevertheless, there are also several limitations in the current work, deserving future investigation. First, this paper only considered monolithic speed distributions, and differences in different time periods within-day are not considered. Second, interaction between factors in the regression analysis were not considered and thus left for further research. In addition, besides data collected by stand-alone GPS devices, other sources of data (e.g., mobile-phone and smart-phone based) should also be included in the future to reach multi-faceted comparisons.

Acknowledgments

This work was supported by a research grant (71371143) from the National Natural Science Foundation of China (NSFC). The work of the first author was supported by Program for Changjiang Scholars and Innovative Research Team in University. This study is partly supported by Dutch Science Foundation (NWO). The authors take sole responsibility for all views and opinions expressed in this paper.

Appendix A

See Tables A1 and A2.

Table A1
Model Parameters of instantaneous speed distributions.

Road Section	Direction	p_0	k_1	λ_1	k_2	λ_2	η
1	N → S	0.0944	1.2142	12.4045	3.6686	33.0861	0.3221
1	S → N	0.5393	0.9051	5.2203	2.1436	21.2234	0.2321
2	N → S	0.1922	1.0815	14.5684	4.3189	34.3989	0.3192
2	S → N	0.3151	1.0014	9.2291	2.8078	22.1782	0.6547
3	N → S	0.2974	1.0749	5.2655	2.7293	32.1020	0.2146
3	S → N	0.2550	1.0795	11.6572	3.8060	37.6088	0.2789
4	N → S	0.3185	0.9167	7.4022	2.5889	23.4529	0.4426
4	S → N	0.4544	1.0150	11.2236	2.9009	21.1254	0.6916
5	N → S	0.2540	1.0027	10.6963	3.2077	28.1307	0.3684
5	S → N	0.1326	1.1281	11.1516	3.1166	33.2343	0.3075
6	N → S	0.3055	1.1763	7.7392	3.0377	28.5693	0.4275
6	S → N	0.2665	1.0286	8.9926	2.4137	23.9480	0.3921
7	N → S	0.5353	0.9872	5.0147	2.8262	27.6988	0.2457
7	S → N	0.2739	1.0500	10.1410	3.0262	31.3112	0.3675
8	N → S	0.1672	0.9560	8.6622	3.1535	33.5363	0.1161
8	S → N	0.4905	1.0503	3.8280	2.3245	26.1556	0.2847
9	N → S	0.4411	0.9733	8.4628	2.4922	25.7803	0.1485
9	S → N	0.3544	0.8947	6.0871	2.4670	24.0359	0.0819
10	N → S	0.1816	1.1548	11.4203	3.0451	30.1140	0.4675
10	S → N	0.4724	1.0158	8.3757	2.3099	24.4775	0.2429
11	N → S	0.3046	1.0761	10.8912	3.2554	33.8291	0.2423
11	S → N	0.1896	1.1507	11.7397	3.3246	33.4151	0.2417
12	N → S	0.3899	0.9694	7.7624	2.5270	23.6231	0.3102
12	S → N	0.2827	1.4698	8.0094	3.7237	33.8677	0.0709
13	N → S	0.0399	0.7916	4.0386	3.1280	33.1056	0.0344
13	S → N	0.3184	1.1315	10.2559	2.7078	28.1423	0.2952
14	N → S	0.2282	1.2145	14.2474	4.3514	38.3556	0.1286
14	S → N	0.1554	1.2975	14.5712	3.4920	29.5472	0.3102
15	N → S	0.3285	1.0680	7.5715	2.5333	26.3667	0.2701
15	S → N	0.2046	0.9273	10.7989	2.3750	22.2913	0.3169

(*continued on next page*)

Table A1 (*continued*)

Road Section	Direction	p_0	k_1	λ_1	k_2	λ_2	η
16	N → S	0.5794	0.9014	3.2401	1.8792	18.8948	0.1768
16	S → N	0.0619	0.9226	15.5780	4.2398	30.1119	0.1214
17	N → S	0.0974	1.0059	3.9231	4.3732	36.3884	0.0417
17	S → N	0.2271	0.9792	7.4856	3.2882	26.2716	0.1988
18	N → S	0.1101	1.2330	2.7415	4.3173	35.7951	0.0197
18	S → N	0.2356	0.8756	6.2344	2.7211	26.1233	0.1840
19	N → S	0.6252	1.0618	9.9838	2.4471	23.6169	0.7976
19	S → N	0.2064	0.9007	3.2805	1.8758	17.6241	0.1605
20	N → S	0.0890	1.0265	8.1738	3.1399	34.1128	0.0661
20	S → N	0.1710	1.0095	11.9255	4.5570	38.8523	0.4487
21	N → S	0.3831	1.0736	9.8228	3.3845	19.4028	0.6459
21	S → N	0.0165	0.4814	3.2942	3.6488	34.0964	0.0232

Table A2
Model parameters of section speed distributions.

Road section	Direction	n_c	μ_1	μ_2	μ_3	μ_4	μ_5	σ_1	σ_2	σ_3	σ_4	σ_5	η_1	η_2	η_3	η_4	η_5
1	N → S	4	16.87	37.25	27.43	50.59	0.00	3.13	9.93	2.06	1.91	0.00	0.13	0.10	0.65	0.12	0.00
1	S → N	5	5.14	13.58	22.23	20.25	41.50	1.21	2.18	1.74	0.91	4.58	0.11	0.23	0.25	0.20	0.21
2	N → S	5	12.49	14.02	24.36	24.48	33.03	0.71	5.10	2.87	1.28	9.13	0.09	0.14	0.56	0.14	0.07
2	S → N	5	10.06	17.67	16.08	29.70	33.17	2.91	1.23	0.55	9.19	2.18	0.33	0.26	0.15	0.06	0.20
3	N → S	3	25.14	31.62	45.64	0.00	0.00	3.03	5.10	5.18	0.00	0.00	0.36	0.55	0.09	0.00	0.00
3	S → N	5	14.92	23.22	27.71	36.98	45.51	1.25	0.58	4.40	3.44	3.94	0.01	0.01	0.55	0.27	0.16
4	N → S	5	14.01	17.89	29.82	25.95	50.98	2.91	1.60	4.53	1.83	4.26	0.16	0.24	0.03	0.55	0.02
4	S → N	5	11.13	15.50	28.64	22.99	46.73	3.75	1.20	3.84	1.71	3.99	0.23	0.27	0.02	0.40	0.09
5	N → S	5	11.42	20.53	16.46	26.92	40.05	1.87	1.50	1.49	2.38	3.88	0.02	0.37	0.10	0.43	0.08
5	S → N	4	17.68	27.22	42.35	38.86	0.00	4.01	2.38	4.59	2.75	0.00	0.33	0.24	0.05	0.38	0.00
6	N → S	4	18.17	24.74	27.54	36.42	0.00	2.98	2.01	10.44	2.89	0.00	0.47	0.29	0.07	0.18	0.00
6	S → N	4	16.44	19.28	25.02	36.20	0.00	3.31	1.22	2.11	4.46	0.00	0.46	0.09	0.29	0.17	0.00
7	N → S	3	22.53	33.48	47.55	0.00	0.00	4.88	2.57	4.58	0.00	0.00	0.57	0.32	0.11	0.00	0.00
7	S → N	4	23.00	24.14	32.47	47.34	0.00	4.87	2.09	3.18	3.44	0.00	0.38	0.26	0.27	0.08	0.00
8	N → S	4	19.59	26.09	31.53	38.60	0.00	1.64	2.51	7.64	2.95	0.00	0.04	0.14	0.34	0.48	0.00
8	S → N	4	18.38	24.77	45.56	35.57	0.00	4.51	1.79	8.82	2.82	0.00	0.38	0.23	0.01	0.38	0.00
9	N → S	5	3.47	16.38	17.64	33.89	32.30	0.68	4.02	0.85	8.78	3.73	0.01	0.37	0.09	0.04	0.49
9	S → N	3	16.06	27.11	30.77	0.00	0.00	1.47	9.42	2.21	0.00	0.00	0.24	0.44	0.32	0.00	0.00
10	N → S	5	16.20	13.06	21.53	33.80	34.78	1.08	1.16	2.97	3.91	11.65	0.18	0.06	0.54	0.20	0.02
10	S → N	5	6.35	18.42	24.30	32.27	52.31	0.74	4.25	1.29	3.57	5.36	0.02	0.67	0.14	0.16	0.00
11	N → S	4	15.66	23.72	28.79	39.68	0.00	3.55	3.26	2.18	4.60	0.00	0.01	0.28	0.25	0.47	0.00
11	S → N	4	15.34	25.40	33.26	44.76	0.00	2.56	3.84	2.99	3.76	0.00	0.02	0.44	0.36	0.18	0.00
12	N → S	4	18.60	24.08	34.59	54.69	0.00	3.92	1.90	3.38	4.52	0.00	0.54	0.29	0.16	0.00	0.00
12	S → N	5	22.43	18.17	25.87	34.33	44.91	6.37	0.77	1.70	3.02	6.89	0.11	0.05	0.24	0.53	0.08
13	N → S	4	16.13	22.54	26.17	44.96	0.00	1.73	1.50	3.47	3.42	0.00	0.07	0.19	0.42	0.32	0.00
13	S → N	3	22.64	27.21	49.10	0.00	0.00	7.14	1.86	3.98	0.00	0.00	0.60	0.31	0.09	0.00	0.00
14	N → S	5	9.46	22.59	29.93	40.10	44.47	0.42	5.21	2.48	4.14	2.07	0.00	0.12	0.18	0.47	0.23
14	S → N	4	21.20	18.77	27.03	30.86	0.00	1.51	2.87	1.67	7.47	0.00	0.23	0.21	0.42	0.14	0.00
15	N → S	3	21.99	29.44	41.23	0.00	0.00	4.76	1.88	4.87	0.00	0.00	0.77	0.15	0.07	0.00	0.00
15	S → N	4	9.82	18.54	26.54	45.36	0.00	1.20	3.39	5.65	6.98	0.00	0.06	0.67	0.19	0.08	0.00
16	N → S	4	11.96	17.84	24.15	56.75	0.00	1.21	7.62	2.41	3.14	0.00	0.31	0.22	0.47	0.00	0.00
16	S → N	1	25.77	0.00	0.00	0.00	0.00	6.81	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
17	N → S	3	21.68	31.41	39.03	0.00	0.00	3.25	2.85	12.20	0.00	0.00	0.07	0.87	0.06	0.00	0.00
17	S → N	4	5.71	22.40	27.02	48.95	0.00	0.58	6.17	1.98	6.61	0.00	0.03	0.70	0.23	0.04	0.00
18	N → S	2	25.32	43.55	0.00	0.00	0.00	5.45	3.59	0.00	0.00	0.00	0.16	0.84	0.00	0.00	0.00
18	S → N	5	6.95	18.42	20.19	33.22	32.45	0.77	4.99	0.90	2.70	9.40	0.11	0.37	0.10	0.16	0.26
19	N → S	5	10.34	16.09	14.19	31.83	26.68	2.74	1.55	0.71	2.88	9.32	0.19	0.27	0.18	0.17	0.20
19	S → N	3	13.29	18.54	30.55	0.00	0.00	1.31	1.84	9.07	0.00	0.00	0.20	0.46	0.34	0.00	0.00
20	N → S	4	26.06	27.59	34.27	42.07	0.00	4.83	1.90	2.49	4.36	0.00	0.20	0.13	0.37	0.30	0.00
20	S → N	3	18.81	30.34	44.12	0.00	0.00	4.47	4.40	4.70	0.00	0.00	0.18	0.38	0.43	0.00	0.00
21	N → S	5	11.70	21.59	17.41	52.83	34.77	1.72	2.48	1.62	4.23	4.13	0.18	0.07	0.50	0.01	0.24
21	S → N	2	20.84	36.49	0.00	0.00	0.00	2.21	6.47	0.00	0.00	0.00	0.09	0.91	0.00	0.00	0.00

References

Brakatsoulas, S., Pfoser, D., Salas, R., Wenk, C., 2005. On Map-Matching Vehicle Tracking Data. In: Proc. 31st Int. Conf. Very Large Data Bases, pp. 853–864.
Chen, Y., Garcia, E.K., Gupta, M.R., et al, 2010. Similarity-based classification: concepts and algorithms. J. Mach. Learn. Res. 10, 747–776.
Cortés, C.E., Gibson, J., Gschwender, A., et al, 2011. Commercial bus speed diagnosis based on GPS-monitored data. Transp. Res. Part C Emerg. Technol. 19, 695–707.

Dey, P.P., Chandra, S., Gangopadhaya, S., 2006. Speed distribution curves under mixed traffic conditions. J. Transp. Eng. 132, 475–481. http://dx.doi.org/10.1061/(ASCE)0733-947X(2006) 132:6(475).
Eichler, M., Daganzo, C.F., 2006. Bus lanes with intermittent priority: Strategy formulae and an evaluation. Transp. Res. Part B Methodol. 40, 731–744.
Endres, D.M., Schindelin, J.E., 2003. A new metric for probability distributions. IEEE Trans. Inf. Theory. http://dx.doi.org/10.1109/TIT.2003.813506.
Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer.
Gerlough, D.L., Huber, M.J., 1975. Traffic flow theory: a monograph Special Report 165. Presented at the Transportation Research Board, National Research Council, Washington, D.C.
Ghilani, C.D., Wolf, P.R., 2007. Elementary Surveying - An Introduction to Geomatics. Pearson Prentice Hall, New Jersey.
Hershey, J.R., Olsen, P.A., 2007. Approximating the Kullback Leibler divergence between Gaussian mixture models. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. http://dx.doi.org/10.1109/ICASSP.2007.366913.
Jun, J., 2010. Understanding the variability of speed distributions under mixed traffic conditions caused by holiday traffic. Transp. Res. Part C Emerg. Technol. 18, 599–610. http://dx.doi.org/10.1016/j.trc.2009.12.005.
Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Biometrics.
Kendall, M.G., Stuart, A., Ord, J.K., 1968. The Advanced Theory of Statistics. Charles Griffin & Co., Ltd., London.
Ko, J., Guensler, R.L., 2005. Characterization of Congestion Based on Speed Distribution: A Statistical Approach Using Gaussian Mixture Model. Presented at Transportation Research Board, National Research Council, Washington, D.C.
Kullback, S., Leibler, R.A., 1951. On Information and Sufficiency. Ann. Math. Stat. 22, 79–86.
Kyriakopoulou, N., Photis, Y.N., Kanaroglou, P., 2016. Mathematical characterization of spatiotemporal congested traffic patterns: mixed speed data analysis in the greater Toronto and Hamilton area. Canada. Transp. Plan. Technol. 39, 1–11.
Leong, H.J.W., 1968. The distribution and trend of free speeds on two-lane two-way rural highways in New South Wales. In: Australian Road Research Board (ARRB) Conference, 4th, 1968, Melbourne.
Liao, F., Arentze, T., Timmermans, H., 2013. Incorporating space-time constraints and activity-travel time profiles in a multi-state supernetwork approach to individual activity-travel scheduling. Transp. Res. Part B Methodol. 55, 41–58.
Liao, F., 2016. Modeling duration choice in space–time multi-state supernetworks for individual activity-travel scheduling. Transp. Res. Part C Emerg. Technol. 69, 16–35.
Lindner, J., 1965. A contribution to the statistical analysis of speed distributions. In: Proceedings of the 3rd International Symposium on the Theory of Traffic Flow, New York.
Llorca, C., Moreno, A.T., Lenorzer, A., Casas, J., Garcia, A., 2015. Development of a new microscopic passing maneuver model for two-lane rural roads. Transp. Res. Part C Emerg. Technol. 52, 157–172.
Ma, X., Wang, Y., 2014. Development of a data-driven platform for transit performance measures using smart card and GPS data. J. Transp. Eng. 140.
Marchal, P., Madre, J.L., Yuan, S., 2011. Postprocessing Procedures for Person-Based GPS Data Collected in the French National Travel Survey 2007–2008. Transportation Research Record Journal of the Transportation Research Board 2246, 47–54.
Maurya, A.K., Dey, S., Das, S., 2015. Speed and Time Headway Distribution under Mixed Traffic Condition. In: Presented at International Conference of Eastern Asia Society for Transportation Studies.
May, A.D., 1990. Traffic Flow Fundamentals. Prentice-Hall, New Jersey.
Mclachlan, G.J., Krishnan, T., 2007. The EM Algorithm and Extensions. John Wiley & Sons Inc, New York.
Mclean, J.R., 1978. Observed speed distributions and rural road traffic operations. In: Proceedings of the 9th Australian Road Research Board Conference, Part 5. Australian Road Research Board, Vermont South, Victoria, Australia.
Park, B.J., Zhang, Y., Lord, D., 2010. Bayesian mixture modeling approach to account for heterogeneity in speed data. Transp. Res. Part B Methodol. 44, 662–673.
Park, B., Schneeberger, J.D., 2003. Microscopic simulation model calibration and validation: case study of VISSIM simulation model for a coordinated actuated signal system. Transp. Res. Rec. 1856, 185–192.
Park, H.S., Jun, C.H., 2009. A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. 36, 3336–3341.
Quddus, M.A., Ochieng, W.Y., Noland, R.B., 2007. Current map-matching algorithms for transport applications: State-of-the art and future research directions. Transp. Res. Part C Emerg. Technol. 15, 312–328.
Rossi, R., Gastaldi, M., Pascucci, F., 2014. Empirical analysis of vehicle time headways and speeds on rural two-lane, two-way roads. Transp. Res. Rec. 2422, 141–149.
Rudin, W., 1973. Functional Analysis. McGraw-Hill Inc, New York.
Szarmes, M., Ryan, S., Lachapelle, G., et al, 1997. DGPS high accuracy aircraft velocity determination using Doppler measurements. In: Proceedings of the International Symposium on Kinematic Systems (KIS97), Banff, Alberta, Canada.
Shen, L., Stopher, P.R., 2014. Review of GPS travel survey and GPS data-processing methods. Transport Rev. 34, 316–334.
Thajchayapong, S., Garcia-Trevino, E.S., Barria, J.A., 2013. Distributed classification of traffic anomalies using microscopic traffic variables. IEEE Trans. Intell. Transp. Syst. 14, 448–458.
Titterington, D.M., Smith, A.F.M., Makov, U.E., 1985. Statistical Analysis of Finite Mixture Distributions. Wiley.
Wang, Y., Dong, W., Zhang, L., Chin, D., Papageorgiou, M., Rose, G., Young, W., 2012. Speed modeling and travel time estimation based on truncated normal and lognormal distributions. Transp. Res. Rec. 2315, 66–72. http://dx.doi.org/10.3141/2315-07.
Whissell, J.S., Clarke, C.L., 2013. A. Effective measures for inter-document similarity. In: ACM International Conference on Conference on Information & Knowledge Management.
Xuan, Y., Argote, J., Daganzo, C.F., 2011. Dynamic bus holding strategies for schedule reliability: optimal linear control and performance analysis. Transp. Res. Part B Methodol. 45, 1831–1845.
Yao, J., Shi, F., An, S., Wang, J., 2015. Evaluation of exclusive bus lanes in a bi-modal degradable road network. Transp. Res. Part C Emerg. Technol. 60, 36–51.
Yu, R., Abdel-Aty, M., 2014. An optimal variable speed limits system to ameliorate traffic safety risk. Transp. Res. Part C Emerg. Technol. 46, 235–246. http://dx.doi.org/10.1016/j.trc.2014.05.016.
Zhang, L., Weng, J., Chen, Z., 2014. Characteristic Analysis of Bus Travel Speed on Commuting Corridors Based on GPS Data. Cota International Conference of Transportation Professionals, pp. 1443–1453.
Zou, Y., Zhang, Y., 2011. Use of skew-normal and skew-t distributions for mixture modeling of freeway speed data. Transp. Res. Rec. 2260, 67–75.