Journal of Southeast University (English Edition) Vol. 28, No. 4, pp. 480 - 484 Dec. 2012 ISSN 1003—7985

Characterizing heterogeneity in vehicular traffic speed using two-step cluster analysis

Pan Yiyong[1] Sun Lu[1,2]

([1]School of Transportation, Southeast University, Nanjing 210096)

([2]Department of Civil Engineering, The Catholic University of America, Washington DC 20064, USA)

Abstract: In order to analyze the heterogeneity in vehicular traffic speed, a new method that integrates cluster analysis and probability distribution function fitting is presented. First, for identifying the optimal number of clusters, the two-step cluster method is applied to analyze actual speed data, which suggests that dividing speed data into two clusters can best reflect the intrinsic patterns of traffic flows. Such information is then taken as guidance in probability distribution function fitting. The normal, skew-normal and skew-t distribution functions are used to fit the probability distribution of each cluster respectively, which suggests that the skew-t distribution has the highest fitting accuracy; the second is skew-normal distribution; the worst is normal distribution. Model analysis results demonstrate that the proposed mixture model has a better fitting and generalization capability than the conventional single model. In addition, the new method is more flexible in terms of data fitting and can provide a more accurate model of speed distribution.

Key words: speed distribution; heterogeneity; mixture model; cluster analysis

doi: 10.3969/j.issn.1003-7985.2012.04.019

Speed is an important measurement of the traffic performance of a highway system[1]. Most analytical and simulation models of traffic use speed as the performance measurement of a transportation system[2]. An appropriate mathematical distribution can help describe speed characteristics and is useful for developing and validating microscopic traffic simulation models. It is necessary to find an appropriate mathematical distribution to describe speed data.

Traditionally, normal, log-normal and other forms of distribution have been used to describe speed data when the characteristics of speed data are more or less homogeneous[3-5]. However, if the characteristics of speed data become heterogeneous and speed distribution exhibits bimodality (or multimodality), the unimodal distribution model fails to obtain a satisfactory fit. Dey et al.[6] used a single normal mixture distribution to represent the bimodal speed distribution under a mixed traffic situation. Ko and Guensler[7] analyzed speed data by assuming that the speed distribution over a given time period has a form of mixed normal distribution. Park et al.[8] captured the heterogeneity in speed data through the finite mixture of normal distributions. The results show that the finite mixture of normal distributions can effectively describe the heterogeneity of speed data. Jun[9] investigated traffic congestion trends by speed patterns during holiday travel periods using the normal mixture model and the expectation maximization (EM) algorithm. Zou et al.[10] proposed the skew-normal and skew-t mixture models to fit speed data to capture excess skewness, kurtosis and bimodality present in speed distribution. However, a major difficulty associated with mixture models is that the number of components in mixture models is determined in an ad hoc and subjective manner without any intelligent judgments and the criteria that are adopted to classify the speed data to capture the heterogeneity of speed data may be arbitrary. This situation becomes more severe when more than two components in the model are considered.

The motivations of this paper are twofold. First, this paper presents a methodological framework that combines the advantages of cluster analysis and probability distribution function fitting to solve the aforementioned difficulties. Secondly, cluster analysis can capture the heterogeneity in speed data and optimally and automatically divide speed data into some components, so the causes of different speed distributions can be identified through investigating the components.

1 Methodologies

1.1 Two-step cluster analysis

The two-step cluster method is developed for the analysis of large data sets. It only requires the data input and can also automatically select the optimal number of clusters. It has two steps: 1) Pre-clustering the cases into many small sub-clusters; 2) Clustering the sub-clusters

Received 2012-07-18.

Biographies: Pan Yiyong (1980—), male, graduate; Sun Lu (corresponding author), male, doctor, professor, sunl@cua.edu.

Foundation items: The National Science Foundation by Changjiang Scholarship of Ministry of Education of China (No. BCS-0527508), the Joint Research Fund for Overseas Natural Science of China (No. 51250110075), the Natural Science Foundation of Jiangsu Province (No. BK200910046), the Postdoctoral Science Foundation of Jiangsu Province (No. 0901005C).

Citation: Pan Yiyong, Sun Lu. Characterizing heterogeneity in vehicular traffic speed using two-step cluster analysis[J]. Journal of Southeast University (English Edition), 2012, 28(4): 480 - 484. [doi: 10.3969/j.issn.1003-7985.2012.04.019]

resulting from the pre-clustering step into the optimal number of clusters.

In the pre-clustering step, it scans the data records one by one and decides whether the current record can be added to one of the previously formed clusters or it starts a new cluster based on the following log-likelihood distance criterion[11],

$$d(i,j) = \xi_i + \xi_j - \xi_{\langle i,j\rangle} \tag{1}$$

where

$$\xi_s = -N_s\left(\sum_{k=1}^{K}\frac{1}{2}\log(\hat{\sigma}_k^2 + \hat{\sigma}_{sk}^2)\right) \qquad s = i, j, \langle i,j\rangle \tag{2}$$

where $d(i,j)$ is the log-likelihood distance between clusters i and j; the $\langle i,j\rangle$ index represents the cluster formed by combining clusters i and j; K is the total number of continuous variables; $\hat{\sigma}_k^2$ is the estimated variance of the continuous variable k for the entire dataset; $\hat{\sigma}_{sk}^2$ is the estimated variance of the continuous variable k in cluster j.

The clustering step takes sub-clusters resulting from the pre-clustering step as input and then groups them into the desired number of clusters. Similar to agglomerative hierarchical clustering[11], those clusters with the minimum distances $d(i,j)$ are merged in each step. The process repeats with a new set of clusters until all the clusters have been merged. Thus, it is quite simple to compare the solutions with a different number of clusters.

The number of clusters can be automatically determined. A two phase estimator is used. In the first stage, the Akaike information criterion (AIC) or the Bayesian information criterion (BIC)[11] for each number of clusters within a specified range according to

$$\left.\begin{aligned}\mathrm{AIC}(J) &= -2\sum_{j=1}^{J}\xi_j + 2m_J\\ \mathrm{BIC}(J) &= -2\sum_{j=1}^{J}\xi_j + m_J\log N\end{aligned}\right\} \tag{3}$$

is computed and used to find a good initial estimate of the maximum number of clusters, where J is the number of clusters, $m_J = 2KJ$. Let $d(J) = \mathrm{BIC}(J) - \mathrm{BIC}(J+1)$. The maximum number of clusters is set equal to the number of clusters when the ratio $d(J)/d(1)$ is smaller than c (currently $c = 0.04$) for the first time. The second stage uses the ratio change $R(k)$ in distance for k clusters, defined as $R(k) = d_{k-1}/d_k$, where d_{k-1} is the distance if k clusters are merged to $k-1$ clusters. The distance d_k is defined similarly. The number of clusters is obtained when a big jump in the ratio change occurs.

1.2 Probability distribution

1.2.1 Normal distribution

The normal distribution of a random variable X is a continuous probability distribution that has a bell-shaped probability density function, known as the Gaussian function[10],

$$f(x\mid\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{4}$$

The parameter μ is the mean or expectation and σ^2 is the variance. σ is known as the standard deviation. The normal distribution is usually denoted by $X \sim N(\mu,\sigma^2)$.

1.2.2 Skew-normal distribution

A random variable X has a skew-normal distribution with location parameter μ, scale parameter σ^2 and skewness parameter λ if its density follows[10] the form:

$$\psi(x\mid\mu,\sigma^2,\lambda) = \frac{2}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\lambda\frac{x-\mu}{\sigma}\right) \tag{5}$$

where $f(\cdot)$ and $\Phi(\cdot)$ are the standard normal density function and the cumulative distribution function, respectively. The skew-normal distribution is usually denoted by $X \sim \mathrm{SN}(\mu,\sigma^2,\lambda)$.

1.2.3 Skew-t distribution

A random variable Y follows a skew-t distribution with location parameter μ, scale parameter σ^2, skewness parameter λ and degrees of freedom ν if it has the following representation[10]:

$$Y = \mu + \sigma\frac{X}{\sqrt{W}} \qquad X \sim \mathrm{SN}(\mu,\sigma^2,\lambda),\ W \sim \Gamma\left(\frac{\nu}{2},\frac{\nu}{2}\right) \tag{6}$$

where $X \sim \mathrm{SN}(\mu,\sigma^2,\lambda)$ is the standard skew normal distribution and independent of the Gamma distribution $\Gamma(\alpha,\beta)$ with mean α/β. The density of Y follows the form[10]:

$$\psi(y\mid\mu,\sigma^2,\lambda,\nu) = \frac{2}{\sigma}t_\nu(x_y)\,T_{\nu+1}\left(\lambda x_y\sqrt{\frac{\nu+1}{\nu+x_y^2}}\right) \tag{7}$$

where $x_y = (y-\mu)/\sigma$, t_ν and T_ν denote the standard Student-t density function and the cumulative function with ν degrees of freedom, respectively. The skew-t distribution is usually denoted as $Y \sim \mathrm{ST}(\mu,\sigma^2,\lambda,\nu)$.

2 Traffic Data Sets

The actual traffic data of roadway: I-10 east bound (sensor ID: L2-0010E-562.581) on July 2, 2001 are collected from TransGuide program[12], the Advanced Traffic Management System (ATMS) at San Antonio, Texas. TransGuide records speed, volume and occupancy from individual lanes of roads at a 20 s interval using loop detectors and video cameras. The dataset contains 4 320 records during a single day. Before the collected traffic data are used for this study, data quality control and preprocessing are conducted to ensure data integrity and correctness. Some of the records show zero speed and some show null vehicle presence. These data are removed from

the datasets. Eventually, the dataset used for investigation contains 3 258 effective records.

The histogram in Fig. 1 displays the frequency of the speed data, which appears to be multimodal with different patterns. Obviously, there are two patterns: the high speed group and the low speed group. Meanwhile, the single normal, skew-normal and skew-t distributions are used to capture the distribution of speed data. The normal, skew-normal and skew-t distribution functions are $N(58.76, 11.87)$, SN(69.37, 40.26, −3.96) and ST(66.56, 36.63, 2.03, 1.83), respectively.

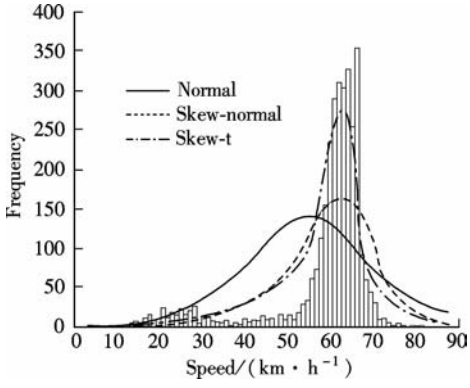

Fig. 1 Histogram and three probability distributions of speed data

To capture the speed data heterogeneity, the two-step cluster method is used to classify the speed data. After data segmentation, normal, skew-normal and skew-t distribution functions can be used to capture the distribution of each speed data subset, respectively.

3 Data Segmentation Results

In this section, the two-step cluster method is applied to classify the speed datasets. The variables entered into the computation are the original speeds. The number of clusters and the data of each cluster are obtained automatically and optimally. Fig. 2 shows the resulting clusters through the two-step cluster method. The speed datasets are divided into two clusters, and the cluster quality is very good based on the silhouette measure of cohesion separation[11]. Tab. 1 shows the results of two clusters.

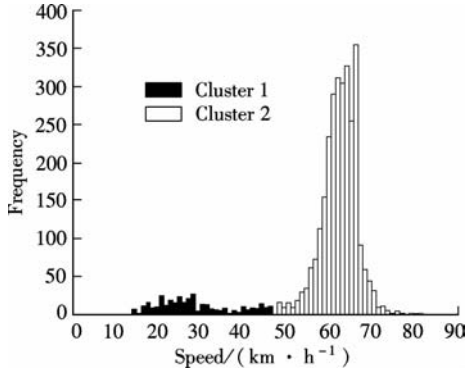

Fig. 2 The resulting clusters of speed dataset

Tab. 1 The results of two clusters

Parameters	Cluster 1	Cluster 2
Proportion/%	11.5	88.5
Mean speed/(km · h^{-1})	28.9	62.64

4 Distribution of Each Cluster

As stated in the previous section, the speed data can be clustered optimally and automatically through the two-step cluster method. The purpose of this section is to develop mixture models of speed distribution and identify heterogeneity in speed data. For the purpose of illustration, normal, skew-normal and skew-t distributions are adopted to capture the distribution of the speed data in each cluster.

Fig. 3 shows the normal distribution of two clusters. All of the estimated normal model parameters are highly statistically significant. The two normal distributions are $N(25.37, 6.187)$ and $N(63.64, 4.189)$. Fig. 4 shows the skew-normal distributions of two clusters. The two skew-normal distributions are SN(21.71, 133.12, 1.3) and SN(65.34, 24.83, −0.92). Fig. 5 shows skew-t distributions of two clusters. The two skew-t distributions are ST(22.56, 124.63, 1.17, 1.83) and ST(64.65, 14.63, −0.73, 5.76).

In this study, only normal distribution, skew-normal distribution and skew-t distribution are chosen to fit the distribution of speed data. The fitting can be further improved if other models of distribution are used. In this

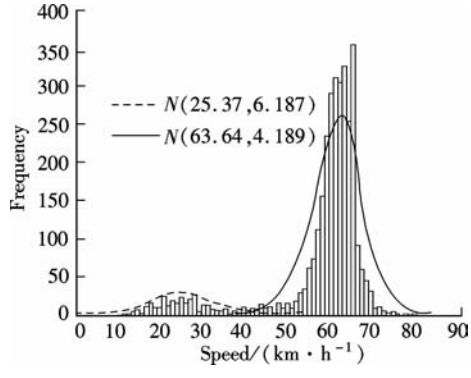

Fig. 3 The normal distribution of two clusters

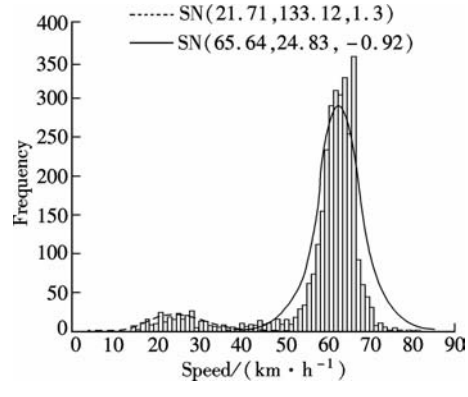

Fig. 4 The skew-normal distribution of two clusters

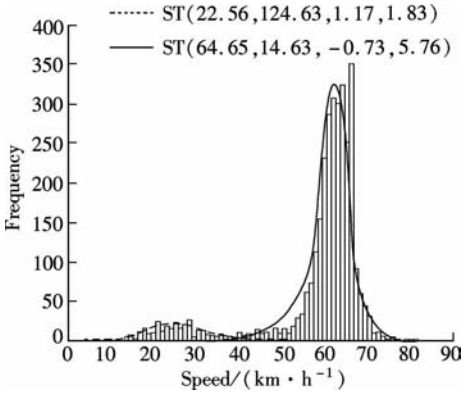

Fig. 5 The skew-t distribution of two clusters

regard, the comparison with other models of distribution can be a subject for further study.

5 Comparison and Verification

To compare the new method that integrates cluster analysis and probability distribution function fitting with the previous method that has a single probability distribution, the fitting error e is defined as [11]

$$e = \frac{\sum_{v=v_{\min}}^{v_{\max}} [p(v) - f(v)]^2}{v_{\max} - v_{\min} + 1} \times 10^5$$

where $p(v)$ is the actual probability density function and $f(v)$ is the theoretical probability density function. $v_{\max}$ and $v_{\min}$ are the maximum speed and minimum speed, respectively. Tab. 2 lists the fitting error of each model.

Tab. 2 The fitting error of each model

Method	Cluster	Normal	Skew-normal	Skew-t
New method	1	95	86	83
	2	68	55	51
	Total	162	141	134
Previous method		547	522	391

As shown in Tab. 2, the fitting error of three models is reduced from 547 to 162, from 522 to 141 and from 391 to 134, respectively. Clearly, the results of the new method that integrates cluster analysis and probability distribution function fitting are more accurate than the previous method that has a single probability distribution.

6 Conclusion

A methodology framework integrating cluster analysis and probability distribution function fitting is presented. The two-step cluster method is applied in cluster analysis to actual speed data for identifying the proper clusters. Normal, skew-normal and skew-t distributions are used to fit the distribution of speed data in each cluster, respectively.

The comparison with the previous method that has a single probability distribution verifies the correctness of the proposed method and demonstrates that the proposed method is superior to the previous one. On the one hand, the new method provides a more practical method because the heterogeneity in speed data can be identified automatically and optimally. On the other hand, the proposed method has better fitting and generalization capability than the previous method. In addition, the new method is more flexible in terms of data fitting and can provide more accurate model of speed distribution. Hence the new method is attractive in dealing with the model of mixed traffic flow.

References

[1] May A D. *Traffic flow fundamentals* [M]. New Jersey: Prentice-Hall, 1990.

[2] Katti B K, Raghavachari S. Modeling of mixed traffic speed data as inputs for the traffic simulation models [M]//*Highway Research Bulletin*. New Delhi, India: Indian Roads Congress, 1986, **28**: 35 - 48.

[3] Leong H J. The distribution and trend of free speeds on two-lane two-way rural highways in New South Wales [C]//*Proceedings of the 4th Australian Road Research Board Conference*. Melbourne, Australia, 1968: 791 - 808.

[4] McLean J R. Observed speed distributions and rural road traffic operations [C]//*Proceedings of the 9th Australian Road Research Board Conference*. Brisbane, Australia, 1978: 235 - 244.

[5] Haight F A, Mosher W W. A practical method for improving the accuracy of vehicular speed distribution measurements [R]. Washington DC, USA: Highway Research Board, 1962: 92 - 116.

[6] Dey P P, Chandra C, Gangopadhaya S. Speed distribution curves under mixed traffic conditions [J]. *Journal of Transportation Engineering*, 2006, **132**(6): 475 - 481.

[7] Ko J, Guensler R L. Characterization of congestion based on speed distribution: a statistical approach using Gaussian mixture model [C/D]//*Transportation Research Board* 2005 *Annual Meeting*. Washington DC, USA: TRB, 2005.

[8] Park B, Zhang Y, Lord D. Bayesian mixture modeling approach to account for heterogeneity in speed data [J]. *Transportation Research Part B*, 2010, **44**(5): 662 - 673.

[9] Jun J. Understanding the variability of speed distributions under mixed traffic conditions caused by holiday traffic [J]. *Transportation Research Part C*, 2010, **18**(4): 599 - 610.

[10] Zou Y J, Zhang Y L. Use of skew-normal and skew-t distributions for mixture modeling of freeway speed data [J]. *Transportation Research Record*, 2011, **2260**: 67 - 75.

[11] SPSS Inc. Two-step cluster analysis [EB/OL]. (2004) [2012-06-20]. http://support.spss.com/tech/stat/Algorithms/12.0/Two-step_cluster.pdf.

[12] Texas Department of Transportation. Roadway network of San Antonio in TransGuide program [EB/OL]. (2006-12-31) [2012-06-20].

基于两步聚类法的交通流速度不均匀性分析

潘义勇[1] 孙 璐[1,2]

([1] 东南大学交通学院, 南京 210096)

([2]Department of Civil Engineering, The Catholic University of America, Washington DC 20064, USA)

摘要:为了分析交通流的速度不均匀性,提出了一种将聚类分析和概率分布函数拟合相结合的新方法. 首先,为了确定最优的子类数,采用两步聚类法对实际的速度数据进行聚类分析,分析表明将速度数据分为2类最能反映交通流的固有类型. 然后,将此信息用于指导概率分布函数拟合,采用正态分布、偏正态分布和偏-T分布函数分别拟合各子类数据的概率分布,发现偏-T分布函数拟合精度最高,偏正态分布次之,正态分布最差. 模型分析结果表明,所提出的混合分布模型比传统单个分布模型具有更好的拟合能力和通用性. 此外,新方法在数据拟合方面更加灵活,且能提供更精确的速度分布模型曲线.

关键词:速度分布;不均匀性;混合模型;聚类分析

中图分类号:U491