Journal of Southeast University (English Edition) Vol. 26, No. 1, pp. 122 – 125 Mar. 2010 ISSN 1003—7985

Spline-based multi-regime traffic stream models

Xiong Wei[1] Sun Lu[2] Zhou Jie[3]

([1] Quality Control Division, Department of Transportation of Anhui Province, Hefei 230051, China)
([2] School of Transportation, Southeast University, Nanjing 210096, China)
([3] Department of Computer Science, Northern Illinois University, DeKalb 60115, USA)

Abstract: In order to develop optimal multi-regime traffic stream models, a new method that integrates cluster analysis and B-spline regression is presented. First, for identifying the proper number of regimes, the *K*-means and the fuzzy *c*-means methods are applied in cluster analysis to actual traffic data, which suggests that dividing the traffic flow into two or three clusters can best reflect intrinsic patterns of traffic flows. Such information is then taken as guidance in spline regression, thus significantly reducing the computational burden of estimating spline models. Spline regression is used to estimate the locations of knots and the coefficients of the model so that the global error can be minimized. Model analysis results demonstrate that the proposed spline models have better fitting and generalization capability than the conventional models. In addition, the new method is more flexible in terms of data fitting and can provide smoother traffic stream models.

Key words: traffic stream; cluster analysis; spline regression; optimization

Multi-regime traffic stream models provide a considerable improvement over single-regime models due to their capability of capturing different patterns of traffic flows. A major difficulty in developing existing multi-regime models is how to determine breakpoints(also known as knots) between regimes. Segmentation of congested and free-flow conditions is exogenously performed based on the subjective judgment of the model developer[1–2]. Such a treatment is empirical and ad-hoc, heavily relying on the modelers' engineering experience. The situation becomes even more severe when more than two regimes are considered since multiple breakpoints need to be determined[3]. Subjective judgments of model developers are not the only issues. Many models have been based on very limited empirical observations, observations in the wrong locations, highly aggregated data, and erroneous data-processing techniques. A traffic stream model should reflect actual measured traffic phenomena that are found to be reproducible. Meanwhile, it is desirable that such a model generates the best fitting to traffic data. This paper addresses the difficulty of breakpoint determination from a data-mining viewpoint. It presents a methodological framework that combines cluster analysis with spline regression to develop optimal traffic stream models.

Received 2009-06-30.
Biographies: Xiong Wei (1971—), male, senior engineer; Sun Lu (corresponding author), male, doctor, professor, sunl@cua.edu.
Foundation item: The US National Science Foundation (No. BCS-0527508).
Citation: Xiong Wei, Sun Lu, Zhou Jie. Spline-based multi-regime traffic stream models[J]. Journal of Southeast University (English Edition), 2010, 26(1): 122 – 125.

1 Methodology

Cluster analysis belongs to unsupervised learning methods and addresses the problem of data segmentation[4]. Cluster analysis divides scattered data into a number of clusters by defining and quantifying dissimilarities between individual data points or patterns[4]. The *K*-means and the fuzzy *c*-means methods are among the most popular methods of cluster analysis. Here, the symbols K and c both stand for the number of clusters in accordance with the custom in pattern recognition terminology.

Our goal of adopting spline regression is to achieve a simple, flexible and accurate model that provides global optimization capability. Spline functions provide a versatile technique for regression and approximation. Splines inherently only require a small number of parameters to provide a maximum flexibility. They have been extensively used in fields such as numeric analysis and computer aided geometric design[5]. Spline regression uses piecewise low order polynomials to fit data. Although any piecewise polynomial function satisfying the definition can be a spline, B-spline basis functions are the most commonly used spline functions.

B-spline basis functions are typically defined recursively. It starts with the definition of the basis function of degree zero,

$$b_{i,0}(\rho)=\begin{cases}1 & \tau_i\leqslant\rho<\tau_{i+1}\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

where ρ is the occupancy. For the B-spline basis function of degree $p\geqslant 1$, we have

$$b_{i,p}=\frac{\rho-\tau_i}{\tau_{i+p}-\tau_i}b_{i,p-1}(\rho)+\frac{\tau_{i+p+1}-\rho}{\tau_{i+p+1}-\tau_{i+1}}b_{i+1,p-1}(\rho) \tag{2}$$

A multi-regime spline traffic stream model is a linear combination of B-spline basis functions,

$$u(\rho)=\sum_{i=-p}^{g}C_i b_{i,k+1}(\rho) \tag{3}$$

where u is the speed, and C_i represent the B-spline coefficients of the model. In a spline multi-regime stream model, parameters that need to be estimated are the degree p of the B-spline, the number and positions of the knots, and the B-spline coefficient C_i. These parameters are estimated in this paper through the minimization of the following least-square criterion,

$$\min\sum_{i=1}^{N}[u_i-u(\rho_i)]^2 \tag{4}$$

where N is the total number of training samples.

2 Traffic Datasets

Actual traffic data on July 2, 2001 are collected from TransGuide program[6], the Advanced Traffic Management System(ATMS) at San Antonio, Texas. TransGuide records speed, volume and occupancy from all the roadway lanes at 20-s intervals using loop detectors and video cameras. Traffic data of a whole day is obtained from the second left lane of three roadways: I-10 east bound(sensor ID: L2-0010E-562.581), US-281 north bound(sensor ID: L2-0281N-143.895), and Loop-1604 west bound(sensor ID: L2-1604W-034.326). Conventional traffic stream models like those reviewed in section 2 use density as an independent variable to characterize macroscopic traffic flow. In this paper we utilize occupancy instead of density, as the latter cannot be directly measured. Since speed, volume and occupancy possess different units, standardization using Eq. (5) is performed prior to cluster analysis so as to make original traffic data dimensionless.

$$x = \frac{x_{\text{original}} - x_{\text{average}}}{x_{\text{std}}} \tag{5}$$

where x_{original} = (speed, volume, occupancy), x_{average} and x_{std} represent original, average and standard deviation of speed, volume and occupancy, respectively. Eq. (5) can also be used to convert a standardized traffic characteristic back into its original counterpart.

3 Number of Regimes

Since the proper number of regimes is yet to be determined, a close examination of the cluster analysis results and traffic flow domain knowledge may provide a rational insight. For this purpose, we apply the *K*-means and the fuzzy *c*-means methods to partition each one of the three standardized traffic datasets into 2, 3, …, 8 clusters, respectively. Traffic characteristics considered in cluster analysis are speed and occupancy. Two criteria are used in this paper for identifying the proper number of regimes: they are the within-cluster dissimilarity measure and the silhouette plot[4, 7-8].

Within-cluster dissimilarity and average silhouette value against the number of clusters are shown in Fig. 1 and Fig. 2, respectively. A general trend of all three traffic datasets is that dissimilarity measures decrease as the number of clusters increase. Using the *K*-means method, the elbow in W_k occurs when the number of clusters is two or three for three roadways. Using the fuzzy *c*-means method, the largest average silhouette width appears when the number of clusters is two or three for three roadways. Since both the clustering methods show similar indications, from the cluster analysis point of view, a plausible number of clusters should be either two or three for any traffic flow under investigation. This is indeed consistent with existing knowledge on daily traffic in which free flow, congestion and transition are among the most commonly observed traffic phenomena.

Fig. 3 shows three different clusters using *K*-means clustering. Based on the above cluster analysis, the model complexity and the domain knowledge of the traffic flow, in this paper we restrict B-spline regression to no more than three regimes. It should be noted that spline regression may be applied to fit stream models with a large number of regimes, though in that situation both the model complexity and the computational burden increase considerably.

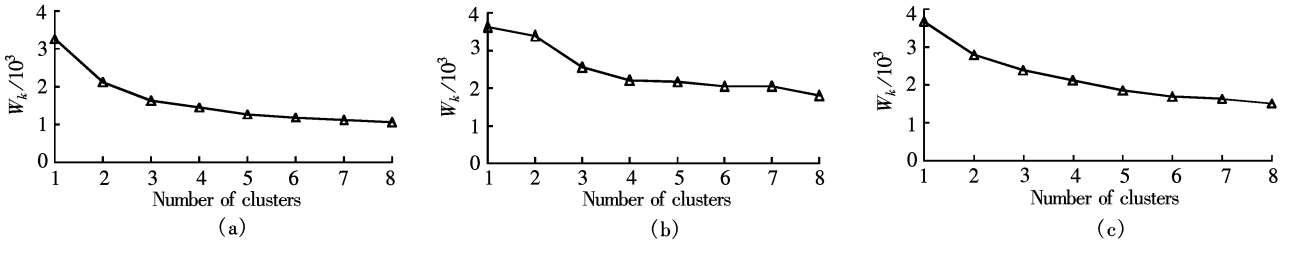

Fig. 1 Within-cluster dissimilarity as a function of number of clusters using *K*-means method. (a) Roadway I-10; (b) Roadway US-281; (c) Roadway Loop-1604

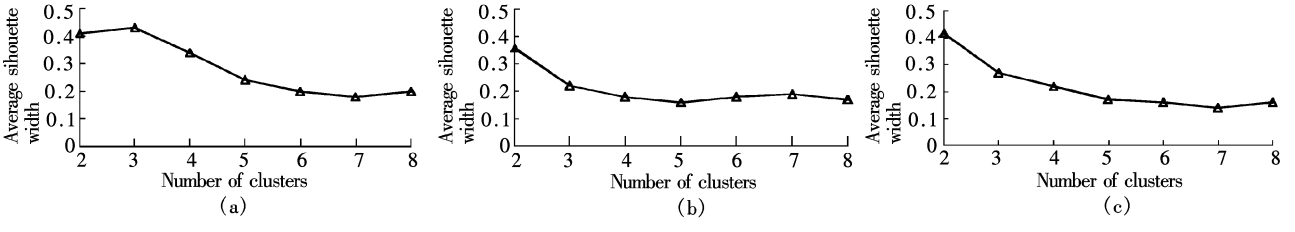

Fig. 2 Average silhouette widths as a function of number of clusters using fuzzy *c*-means method. (a) Roadway I-10; (b) Roadway US-281; (c) Roadway Loop-1604

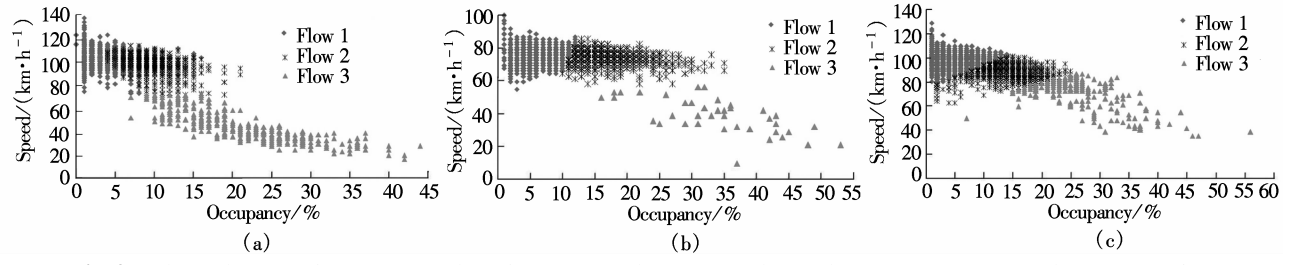

Fig. 3 Three clusters using *K*-means clustering. (a) Roadway I-10; (b) Roadway US-281; (c) Roadway Loop-1604

4 B-Spline-Based Multi-Regime Stream Models

To examine the performance of spline regression models, multi-regime stream models developed using quadratic B-spline and conventional methods are compared with each other. Knot positions in conventional methods are subjectively determined. In developing conventional models, the number of regimes is restricted to no more than three so that it is comparable to spline regression. In addition, only linear functions are used in conventional methods for representing each regime of speed-density relation. The split-sample test is used to examine the performance of spline and conventional traffic stream models[8]. Use traffic data from highway I-10 as an example. One-, two- and three-regime of piecewise linear traffic stream models and quadratic B-spline models are respectively used to fit 70% of the observations. The one corresponding to the minimum least square error is chosen as the best model in respective conventional methods and quadratic B-spline methods.

Fig. 4 provides spline and conventional piecewise linear stream models against raw data for roadways I-10, US-281 and Loop-1604, respectively. Tab. 1 presents model parameters and knots of multi-regime stream models. Tab. 2 lists least square error for training and prediction using spline and conventional three-regime models. Spline models have smaller training and prediction errors than the conventional counterparts, suggesting spline models have better fitting and generalization capability. This is because a global optimization criterion is set forth in spline regression to seek the best knot (break point) positions. On the contrary, the knot position is subjectively determined in conventional methods, which is unlikely to be optimal. Other benefits of spline regression include that they are flexible in terms of data fitting and continuous at knots, providing smoother traffic stream models.

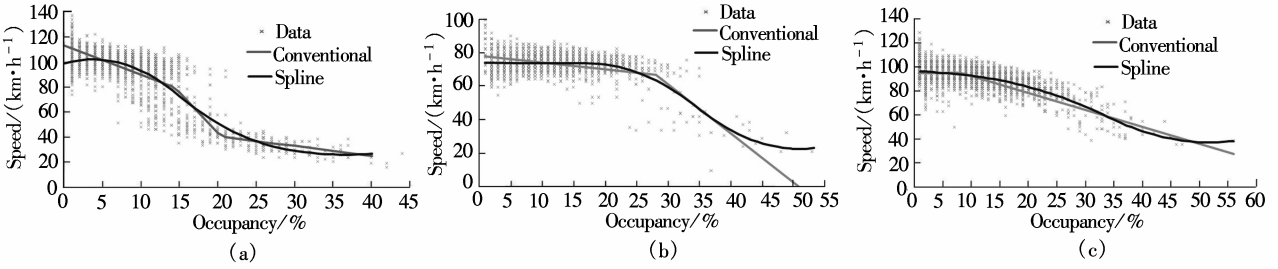

Fig. 4 Spline and conventional multi-regime speed-occupancy models. (a) Roadway I-10; (b) Roadway US-281; (c) Roadway Loop-1604

Tab. 1 Spline and conventional multi-regime stream models

Methods	Roadways	Multi-regime stream models
Conventional piecewise regression	I-10	$u=\begin{cases}71.38-0.147\rho & \rho\leqslant 14.44\\ 109.92-4.14\rho & 14.44\leqslant\rho\leqslant 20.40\\ 35.73-0.50\rho & \rho\geqslant 20.40\end{cases}$
	US-281	$u=\begin{cases}49.26-0.26\rho & \rho\leqslant 28.28\\ 95.25-1.88\rho & \rho\geqslant 28.28\end{cases}$
	Loop-1604	$u=\begin{cases}59.35-0.03\rho & \rho\leqslant 8.29\\ 66.38-0.88\rho & \rho\geqslant 8.29\end{cases}$
Quadratic B-spline regression	I-10	Interior knots: [14.91, 27.75] Coefficients: [62.510, 70.655, 25.567, 14.991, 16.831]
	US-281	Interior knots: [17.26, 34.64] Coefficients: [46.758, 46.612, 46.447, 11.881, 14.372]
	Loop-1604	Interior knots: [12.49 34.24] Coefficients: [60.081, 59.533, 51.900, 20.277, 23.877]

Tab. 2 Comparison of least square error for spline and conventional stream models

Roadways	Spline models/$(km\cdot h^{-1})^2$		Conventional models/$(km\cdot h^{-1})^2$	
	Training	Prediction	Training	Prediction
I-10	178 097	93 472	209 728	104 643
US-281	54 240	25 293	63 329	30 145
Loop-1604	105 380	47 809	120 909	52 202

Note: 70% dataset for training and 30% dataset for prediction

5 Conclusion

A methodological framework for developing multi-regime traffic stream models using cluster analysis and B-spline regression is presented. The new method is a data-driven approach, which does not presume any linearity and monotone at any regime, and non-linearity is automatically taken into account. The benefits of the new method are as follows. First, model parameters include coefficients and knots are globally and optimally determined to produce the best fitting of actual speed-occupancy observations under a specified number of regimes and the order of splines. Secondly, derivative continuity up to one order lower than the highest spline degree can be preserved, which may be desirable in some applications. Thirdly, the new method has great flexibility, as the number of regimes and order of spline can be adjusted to provide the best fitting to actual observations. As a result, nonlinear relationships can be naturally captured by B-splines. This method can also be used for developing multi-regime speed-density relationships.

References

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

[2] Kockelman K M. Modeling traffic's flow-density relation: accommodation of multiple flow regimes and traveler types [J]. *Transportation*, 2001, **28**(4): 363 – 374.

[3] Sun L, Zhou J. Developing multi-regime speed-density relationships using cluster analysis [J]. *Transportation Research Record*, 2005(1934): 64 – 71.

[4] Duda R O, Hart P E, Stork D G. *Pattern classification* [M]. 2nd ed. New York: John Wiley & Sons, Inc, 2001.

[5] Dierckx P. *Curve and surface fitting with splines* [M]. Oxford Science Publications, 1993.

[6] Texas Department of Transportation. Roadway network of San Antonio in TransGuide Program [EB/OL]. (2006-12-31) [2009-06-20]. http://www.transguide.dot.state.tx.us/.

[7] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis [J]. *Journal of Computational and Applied Mathematics*, 1987, **20**(1): 53 – 65.

[8] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning* [M]. New York: Springer, 2001.

基于样条曲线的多域交通流模型

熊 伟[1] 孙 璐[2] 周 洁[3]

([1] 安徽省交通厅质量监督站, 合肥 230051)

([2] 东南大学交通学院, 南京 210096)

([3] 北伊利诺斯州大学计算机科学系, DeKalb 60115)

摘要: 为了建立最优化多域交通流模型, 提出了一种将聚类分析和 B-样条回归分析相结合的新方法. 首先, 为了确定合适的域数, 采用 *K*-means 和模糊 *c*-means 算法对实际交通流数据进行了聚类分析, 分析表明将交通流数据分为 2 类或 3 类最能反映交通流的固有类型. 然后, 将此信息用于指导 B-样条回归分析,可显著减少样条模型参数估计的计算量. 用样条回归分析来估计结点位置和模型参数可使得整体误差最小. 模型分析结果表明, 提出的样条曲线模型比传统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, 且能提供更加光滑的交通流模型曲线.

关键词: 交通流; 聚类分析; 样条回归; 优化

中图分类号: U491