Characterization of Congestion Based on Speed Distribution: A Statistical Approach Using Gaussian Mixture Model

Joonho Ko
Graduate Research Assistant
School of Civil & Environmental Engineering
Georgia Institute of Technology
790 Atlantic Drive, Atlanta, GA 30332-0355
404-385-2376
404-385-2376 (Fax)
Joonho.ko@ce.gatech.edu

Randall L. Guensler, Ph.D.
Associate Professor
School of Civil & Environmental Engineering
Georgia Institute of Technology
790 Atlantic Drive, Atlanta, GA 30332-0355
404-894-0405
404-894-2278 (Fax)
Randall.guensler@ce.gatech.edu

Submitted to: TRB Committee on Highway Capacity and Quality of Service (AHB40) in Division A
Date of submittal: August 1, 2004
Word count: 5,405 words (3,155 words in the manuscript + 2,250 words for figures and tables)

ABSTRACT

Traffic congestion has been one of major problems in modern cities and thus numerous measures have been taken to mitigate it. An appropriate selection of congestion mitigation strategies comes from the sound understanding of the congestion characteristics. Many researchers, thus, have tried to identify the effective ways for measuring level of congestion or roadway performance. As a result of these efforts, various approaches were developed for some specific purposes. It is, however, still necessary to develop new approaches to measuring congestion since congestion cannot be defined in a unique manner and there are numerous situations in which different types of congestion measure are required. This paper proposes a new approach to characterizing the congestion using Gaussian mixture model. This approach assumes that the speed distribution over a given time period has a form of mixed distribution which includes congested and uncongested ones. In addition, it supposes both the speed distributions are normal. Based on these assumptions, it was possible to easily identify the congestion characteristics by exploring the parameters of the distributions. Authors proposed the resulting parameters, means, variances and mixture weights are potential quantifying measures for the severity, variability and duration of congestion, respectively. In particular, this approach is easy to apply since the longitudinal speed data are the only input for the analysis. From the case study using one-week four-freeway segments data, authors concluded that this approach might become a promising one in analyzing congestion characteristics, complementing the existing approaches.

INTRODUCTION

Background

Traffic congestion has been one of major issues that most metropolises are facing and thus, many measures have been taken in order to mitigate congestion. It is believed that identification of congestion characteristics is the first step for such efforts since it is an essential guidance for selecting appropriate measures. Many researchers, thus, have tried to identify the intensity, duration and extent of the congestion using a variety of approaches. Stathopoulos and Karlaftis (*1*), for instance, tried to probabilistically model the duration of traffic congestion using loglogistic function. Thurgood (*2*) developed an index called Freeway Congestion Index, which simultaneously captures the extent and duration of congestion on freeways. In addition, some recent studies have introduced a reliability measure, so called "buffer time index", which shows the effect of congestion on the reliability of travel rates along the roadway (*3-4*). Cottrell (*5*) developed logistic regression models with explanatory variables, AADT/capacity, *K*-factor (i.e., the ratio of the 30th highest hourly volume of the year to the AADT) to predict the occurrence of congestion. Vaziri (*6*) and Hamad and Kikuchi (*7*) applied fuzzy set theory, where multiple congestion measures were combined to get a single comprehensive measure.

It should be admitted that the selection of adequate measures of traffic system performance is not an easy task. This is because congestion is rather subjective and at the same time a location-, facility- and time-dependent matter. Thus, congestion has not yet been clearly defined in a single manner. The congestion, however, can be generally defined as travel time or delay in excess of an agreed-upon norm. The agreed-upon norm may vary by type of transportation facility, geographic location, and time of day (*8*). Many other researchers have also pointed out the aforementioned aspects of congestion and proposed new approaches to quantifying congestion considering those factors.

Congestion index should be developed and selected such that they can effectively and efficiently describe actual traffic conditions and contain useful information. Also, they must be understandable and acceptable to travelers and transportation experts and be well fitted to given purposes. To meet these conditions, the measures require appropriate data in terms of type (e.g. travel time, speed, and traffic volume), temporal, and spatial coverage, etc. Among others, travel time and speed based measures have been most widely used in previous research (*8, 9*). In fact, the measures associated with the time or speed are easy to understand and interpret. The current approaches to measuring roadway performance are well summarized in Medley and Demetsky (*3*).

The target of congestion measures can be an area-wide network, corridors, or link segments. The different type of target requires different data collection efforts. Recently, Intelligent Transportation System (ITS) or Advanced Traffic Management System (ATMS) data are popular resources for transportation research since they are readily available and huge in quantity. Medley and Demetsky (*3*) is an example in which ITS information data were used to develop the corridor-level performance measures. Also, the GPS data from instrumented vehicles are growing as a popular data source for transportation studies. D'Este et al. (*10*) discussed the usefulness of GPS data when developing congestion indices.

Study Purpose

One can expect that for a roadway segment in urban area, longitudinal travel speed data over a day are composed of two main components – congested and uncongested ones. This is seemingly compatible with the fact that one divides urban traffic conditions into peak and off-peak conditions. In this sense, it can be reasonably assumed that a distribution of one-day speed data has a characteristic of bi-modality. And thus, it is expected that parameters resulting from the bi-modality analysis contribute to characterizing congestion for the given roadway section, only if the parameters can be adequately interpreted. Behind of this notion, it is supposed that the speed distribution itself reveals the traffic characteristics of the given location without additional efforts to take into account of other factors such as roadway capacity and free flow speed. The capacity or free flow speed is usually referenced as a base condition of the studied roadway to measure the level of congestion.

The aforementioned approach requires continuous data in terms of time. The traffic data collected by the traffic surveillance cameras as part of ATMS are appropriate data sources for this, because they provide vast sets of continuous traffic data such as average vehicle speed and traffic volume with spatial information. Fortunately, these data are readily available and make the speed distribution based approach suggested in this paper applicable.

The purpose of this paper is to examine the applicability of the congestion characterizing approach based on the assumption of mixed speed distribution using the ATMS data. The suggested approach would be advantageous in that it handles a large amount of data effectively using an established statistical method, and thus it can draw features of the roadway performance from a different perspective not used in existing approaches. In the following sections, the methodology for analyzing speed distribution will be introduced. And then results of a case study will be presented.

METHODOLOGY

Gaussian Mixture Model and EM Algorithm

In some situations, the obtained data have the form of multi-modality in terms of its distribution. In this case, the Gaussian mixture model is a very useful tool in estimating the population density. The Gaussian mixture model is given by the weighted sum of Gaussian densities, $\phi(x;\phi_j)$ with parameters of mean m_j and covariance matrix S_j. In other words, $\theta_j = (m_j,\ S_j)$.

$$f(x) = \sum_{j=1}^{k} \pi_j \phi(x;\theta_j)$$

where, $\phi(x;\phi_j) \sim N(m_j, \sigma_j^2)$ and $\pi_1 + \pi_2 + \cdots + \pi_k = 1,\ \pi_j \geq 0$.

For the learning problem, let's assume a training set $\{x_1, \ldots, x_n\}$, which is composed of independent and identically distributed points sampled from the mixture, and then the task is to estimate the parameters (π_j, m_j, S_j) of the k-components that maximize the log-likelihood $L_k = \sum_{i=1}^{n} \log f_k(x_i)$ (*11*). Direct maximization of L_k is quite difficult numerically, because of the sum of terms inside the logarithm (*12*). The log-likelihood maximization, however, can be simply carried out by the EM (Expectation – Maximization) algorithm using the following iterative update equations for each component j, $j = 1,K ,k$ as represented in Vlassis and Liksa (*11*) and Hastie et al. (*12*).

$$P(j \mid x_i) = \frac{\pi_j \phi(x_i, \theta_j)}{f_k(x_i)},$$

$$\pi_j \equiv \frac{1}{n} \sum_{i=1}^{n} P(j \mid x_i),$$

$$m_j \equiv \frac{\sum_{i=1}^{n} P(j \mid x_i) x_i}{\sum_{i=1}^{n} P(j \mid x_i)}$$

$$S_j \equiv \frac{\sum_{i=1}^{n} P(j \mid x_i)(x_i - m_j)(x_i - m_j)^T}{\sum_{i=1}^{n} P(j \mid x_i)}$$

The iteration should continue until convergence of log-likelihood. For initial guesses for π_j, m_j and S_j, *K*-means clustering analysis can be implemented, while Hastie et al. (*12*) suggests a simpler approach where π_j can be started at the value 0.5 and m_j be selected at random. And S_j can be set equal to the overall sample variance in the simpler approach. It was showed in Hastie et al. (*12*) that EM iteration never decreases the log-likelihood. This Gaussian mixture model was applied to identifying the daily mixed speed distributions, where it was assumed that the speed distributions are normally distributed.

Gaussian Kernel Density Estimation

Before applying the Gaussian mixture model, it is necessary to investigate the form of speed distributions. Although histograms can describe it, sometimes it is hard to judge due to its irregular and bumpy patterns. In this paper, kernel density estimation technique was applied in order to check the form of speed distributions.

Kernel density estimation is an unsupervised learning procedure, which historically precedes kernel regression (*12*). Let's suppose that *N* samples $x_1, \ldots, x_N$, are drawn from a probability density $f_x(x)$, and we try to estimate f_x at a point x_0. A natural local estimate has the form

$$\hat{f}_X(x_0) = \frac{\# x_i \in Neighbor(x_0)}{N\lambda},$$

where *Neighbor(x_0)* is a small metric neighborhood around x_0 of width λ. This estimate is likely to be bumpy. Instead, we can estimate the density using the smooth Parzen estimate

$$\hat{f}_X(x_0) = \frac{1}{N\lambda}\sum_{i=1}^{N} K_\lambda(x_0, x_i) ,$$

where λ is the bandwidth of the kernel function K_λ. For the kernel function, the Gaussian kernel, $K_\lambda(x_0, x_i) = \phi(| x - x_0 | / \lambda)$ is widely used.

The Parzen density estimate is the equivalent of the local average, and improvements have been proposed along the lines of local regression. In $\mathbf{R}^p$ the natural generalization of the Gaussian density estimate follows

$$\hat{f}_X(x_0) = \frac{1}{N(2\lambda^2\pi)^{\frac{p}{2}}}\sum_{i=1}^{N} e^{-\frac{1}{2}(\|x_i - x_0\|/\lambda)^2} .$$

In this study, this Gaussian density estimation technique was used for preliminary speed density analysis.

CASE STUDY

Data

The Georgia Department of Transportation operates the ATMS to tackle the traffic congestion and minimize the impacts of potential incidents on the major roadways. One of the major roles of this system is to provide drivers with real time travel information. To this, a traffic monitoring system has been built with the help of video detectors along major highways as shown in Figure 1. The camera has a capability of image processing which can calculate the average speeds, traffic volume figures, etc.

The case study employed time series of vehicle average speed from the system, where original data were processed to remedy some problematic detector readings like missing values and stuck detector values (*13*). In particular, four locations of the interstate 75 in Atlanta, Georgia were sampled: 1) Northbound I-75/85 near Spring Street, 2) Northbound I-75 near Northside Drive, 3) Southbound I-75 near Collier Road, and 4) Southbound I-75/85 near North Avenue. Weekend data were eliminated from the data set to capture only weekday traffic patterns. In addition, night-time data were not used since vehicle speeds during night times are extremely high and thus not crucial for the traffic management. The average vehicle speeds were measured in 15-minute period. The data features are summarized as following:

- Data collection dates: 8/29/02 (Thursday), 8/30/02 (Friday), 9/2/02 (Monday), 9/3/02 (Tuesday), 9/4/02 (Wednesday)
- Data resolution: 15-min period
- Time period: 06:00 ~ 20:00

Figure 1 shows the four locations where the vehicle speed data were collected along the interstate 75 and 75/85 downtown connector. Two sites on I-75 near Collier Road and Northside Drive are located in urban area, while the others are located in downtown Atlanta. It is expected that such two different types of location should affect the vehicle speed distributions on the time of day basis, and thus different congestion characteristics would be revealed.

Figures 2 and 3 represent vehicle speed patterns over the data collection period. It should be noted that these time series are not continuous since night-time (8 pm ~ 6 am) data are not included in these series. In figure 2, day-to-day speed variations are observed. For example, traffic condition of the site near North Avenue is fairly good on Monday, while the other days show extremely degraded traffic conditions in the afternoon. It is expected that the approach using multi-day speed data can reflect the day-to-day variations in the model and thus provide a more general picture about the performance level of the studied roadway.

It is evident that each location has its own traffic pattern. For example, the location of I-75 southbound near Collier Road contains the morning congestion while the locations of I-75/85 northbound near Spring Street and I-75/85 southbound near North Avenue show afternoon peak. Obviously, this is the typical urban traffic pattern affected by the location characteristics. The examination of the plots reveals that there might be two different speed distributions formed by the congested and normal speeds. In the following sections, those two distributions will be explored.

Preliminary Analysis

As a preliminary analysis, the Gaussian kernel density estimation technique was applied to the speed data. In this step, one can make sure the applicability of mixture model by examining the shapes of estimated probability density functions. Figure 4 illustrates histograms and their corresponding estimated Gaussian kernel density functions for the speed data of sampled locations. It looks that two locations near Spring Street and near North Avenue clearly have a mixed distribution, although the speed distribution of normal traffic condition is dominant. On the contrary, it is hard to see two different distributions in the locations near Northside drive and near Collier Road. One should note, however, there are several observations in the lower speed range. Although their portion is so small, they can make the bimodality assumption still valid.

It is worthwhile to note the distribution of lower speed range since it is possibly due to congestion which is the major target for this research effort. In particular, the speed distributions sampled from downtown area (I-75/85 near Spring Street and North Avenue) clearly illustrate the mixed distributions. This nonparametric density estimation results, unfortunately, do not directly provide measurable estimates such as means of two speed distributions and weights of the mixture.

Gaussian Mixture Analysis

In the previous section, it was identified that speed distributions during a week are mainly composed of two elements. According to this fact, two distributions were estimated based on two-component Gaussian mixture model. Authors assumed that those distributions are normally distributed, and then applied EM algorithm to elicit them. The estimated two-component Gaussian mixtures are visualized in Figure 5 with the estimated parameters. As a whole, the shapes of the curves are very close to those of Gaussian Kernel density estimation. The resulted plot from the I-75/85 Southbound near North Avenue shows a clearer bi-modal feature.

The results of Gaussian mixture model provided the estimated means, variances and weights for the mixed speed distributions. Table 1 summarizes the estimated parameters of two mixtures with those of overall data. Although the overall mean and variance of speed by location can give insights about the roadway performance, it may fail to reveal some aspects about the congestion experienced by the roadway. In this sense, separate examinations of mixed distributions can be a remedial approach for that. Figures 6, 7, and 8 visualize the comparisons, where Gaussian 1 and Gaussian 2 represent the lower (congested traffic condition) and higher (normal traffic condition) speed distributions, respectively.

It is found that speed distributions of Northbound I-75 near Northside Drive and Southbound I-75 Near Collier Road have a dominant distribution that comes from normal traffic conditions, as suggested by the mixture weights of larger than 0.9. Based on this, the traffic conditions of these road sections are seemingly pretty good over the studied time period. One the other hand, the speed distribution of Southbound I-75/85 near North Avenue is rather evenly split into two, where the mixture weights are 0.51 and 0.49 for lower and higher speed distributions, respectively. This implies that this road section experienced relatively severe traffic congestion in terms of duration.

When it comes to means of speed distributions, the location near Northside Drive shows the most uncongested characteristics, since means of two mixed speed distributions are higher than those of other locations. Also, variances of the distributions in this location are the smallest. These facts imply that traffic flow of this location is good and stable, which is also observed in figures 2 and 3. On the contrary, two downtown sites, northbound I-75/85 near Spring Street and southbound I-75/85 near North Avenue, have relatively low means and large variances, which imply bad and unstable traffic conditions in the locations.

It is observed that the estimated model parameters, mean, variance, and mixture weight, represent the characteristics of the traffic conditions in a quantitative manner. Each parameter has its own implication. The mean of lower speed distribution approximates the severity of the congestion. And the mean of higher speed distribution may be used to define the acceptable speed range of the given roadway section. The acceptable speed in this context is the speed that can be most frequently observed during the normal traffic conditions, probably during off-peak. It should be a location specific value. Based on this notion, the acceptable speed for the location near Northside Drive would be around 65 mph, while other locations be about 55 mph. The variance seemingly represents the reliability of the roadway performance, where a higher variance means the instability of traffic condition. The mixture weight can be employed as a congestion index on a scale of 0 to 1, in particular explaining the congestion duration. Of the two weights from lower and higher speed distributions, any weight can be adopted as an index, since their sum is equal to one. When the congestion index is defined by the weight of the Gaussian 1 (lower speed distribution), for example, the lower values indicate the better traffic condition. According to this criterion, the roadway section near Northside Drive has an index of 0.01, which is the best traffic condition, while the congestion index of roadway section near North Avenue is 0.51, which is the worst among four case study sites. It should be noted that although one can interpret the parameters separately, combining them may give comprehensive insights about the congestion.

CONCLUSIONS

This paper proposed an approach to identifying the congestion characteristics of given roadway sections based on the speed distribution. In particular, authors applied the Gaussian Kernel density estimation technique for preliminary data analysis and two-component Gaussian mixture model for obtaining quantitative congestion measures. A case study was carried out using one-week 15-min speed data obtained from four different sites along interstate 75 in Atlanta, Georgia. It was identified that the Gaussian mixture model can be a good tool for describing the characteristics of congestion via the estimated parameters, mean, variance, and mixture weight. It was suggested that the weight estimated from the Gaussian mixture model could be used as a congestion measure representing the congestion duration. In addition, authors proposed that the means and variances might be used as measures for severity and variability of congestion. The same approach can be also applied to other traffic data available such as traffic volume.

In the suggested approach, the characteristics of congestion are modeled based on the bimodal speed distribution without considering other factors like roadway capacity or free-flow speed. The case study, in fact, used only 15-min speed data collected over 5 weekdays. This implies that the proposed approach is relatively easy to apply because data requirement is not demanding, while some congestion measures need estimation of the capacity or free flow speed of the studied roadway section, which is not always an easy task. Behind of this advantage, it is assumed that the speed distribution should reveal the typical traffic conditions of the roadway section. This assumption can be satisfied by appropriately selecting and processing traffic data. Considering that existing approaches should be also based on typical traffic conditions, however, this requirement is not only confined to the approach suggested in this paper.

As a final remark, it should be admitted that there are some limitations in this approach. First, the underlying assumption, the Gaussian mixture, should be justified although the normal distribution is most commonly adopted. The normality requirement might be achieved by enhancing the data size, which depends on the length of analysis period and data resolution. In future research, this issue will be explored in detail. Second, a bi-modality of the speed distribution cannot be always applied although the case study assumes it. Some roadway sections may have more than two modes, e.g., uncongested speed range, interim speed range which lies between uncongested and congested conditions, and congested speed range. This is the reason for implementation of preliminary analysis using the Gaussian kernel density estimation technique before proceeding into the Gaussian mixture model. But the Gaussian Mixture model is still applicable even when n (>2) modes are detected, since two-component mixture model can be easily extended to *n*-component model. In this situation, the interpretation of resulting parameters is the same as suggested in this paper.

REFERENCES

1. Stathopoulos, A., and Karlaftis, M. G. Modeling the Duration of Urban Traffic Congestion, Presented at 81st Annual meeting of the Transportation Research Board, Washington D.C. 2002.
2. Thurgood, Glen S. Development of a Freeway Congestion Index Using an Instrumented Vehicle. *In Transportation Research Record 1494*, TRB, National Research Council, Washington, D.C., 1995, pp. 21-29.
3. Medley, S. B., and Demetsky, M. J., *Development of congestion performance measures using ITS information*, Virginia Transportation Research Council, VTRC 03-R1, 2003.
4. Lomax, T., Turner, S., and Margiotta, R. *Monitoring Urban Roadways in 2001: Examining Reliability and Mobility with Archived Data.* Publication FHWA-OP-03-141. FHWA, U.S. Department of Transportation, 2003.
5. Cottrell, W.D. Estimating the Probability of Freeway Congestion Recurrence*. In Transportation Research Record 1634*. Transportation Research Board, Washington, D.C., 1998, pp. 19-27.
6. Vaziri, M. Development of Highway Congestion Index Using Fuzzy Set Models, *In Transportation Research Record 1802*. Transportation Research Board, Washington, D.C., 2002, pp. 16-22.
7. Hamad, K., and Kikuchi, S. Developing A Measure of Traffic Congestion: A Fuzzy Inference Approach, *In Transportation Research Record 1802*. Transportation Research Board, Washington, D.C., 2002, pp. 77-85.
8. Turner, S. M., Lomax, T., and Levinson, H. Measuring and Estimating Congestion Using Travel Time-Based Procedures. *In Transportation Research Record 156*4, TRB, National Research Council, Washington, D.C., 1996, pp. 11-19.
9. Levinson, H., and T. Lomax. Developing a travel time congestion index. In *Transportation Research Record 1564,* TRB, National Research Council, 1996, pp. 1-10.
10. D'Este, G., R. Zito, and M. Taylor. Using GPS to Measure Traffic System Performance. *Computer-Aided Civil and Infrastructure Engineerin*g, vol. 14, 1999, pp. 255-265.
11. Vlassis, N., and A. Likas. *A greedy EM algorithm for Gaussian mixture learning*. *Neural Processing Letters*, 15(1):77-87, February 2002
12. Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2001.
13. Guensler, R. and M. Oliveira. Vehicle Activity Data Sources and Emissions Estimates for the Jefferson Street Particulate Matter Monitoring Site, Trans/AQ, Inc., 2001.

LIST OF TABLES AND FIGURES

TABLE 1 Comparison of speed distributions

		Northbound I-75/85 Near Spring Street	Northbound I-75 Near Northside Dr.	Southbound I-75 Near Collier Rd.	Southbound I-75/85 Near North Ave.
Overall	Mean	52.5	67.4	52.6	49.5
	Variance	80.2	7.6	18.0	169.3
Gaussian 1 (congested)	Mean	34.5	47.2	42.4	39.9
	Variance	139.6	3.3	72.2	165.5
	Weight	0.15	0.01	0.08	0.51
Gaussian 2 (normal)	Mean	55.5	67.7	53.5	58.4
	Variance	5.8	1.6	3.3	5.5
	Weight	0.85	0.99	0.92	0.49

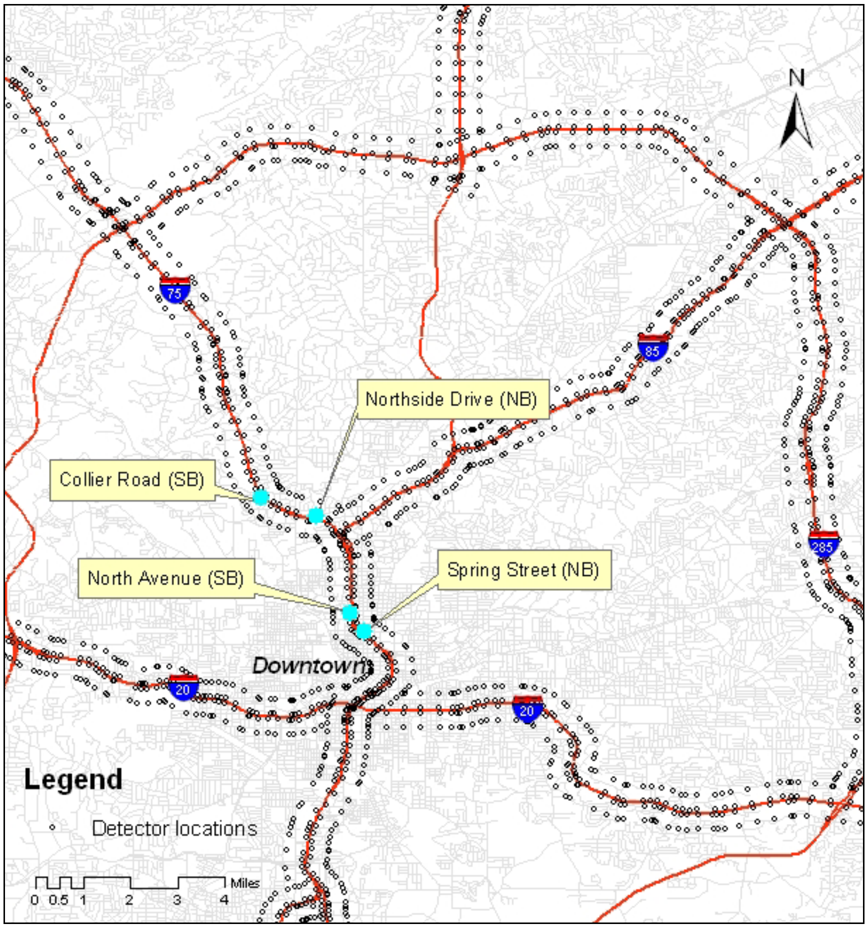


FIGURE 1 Video detector locations

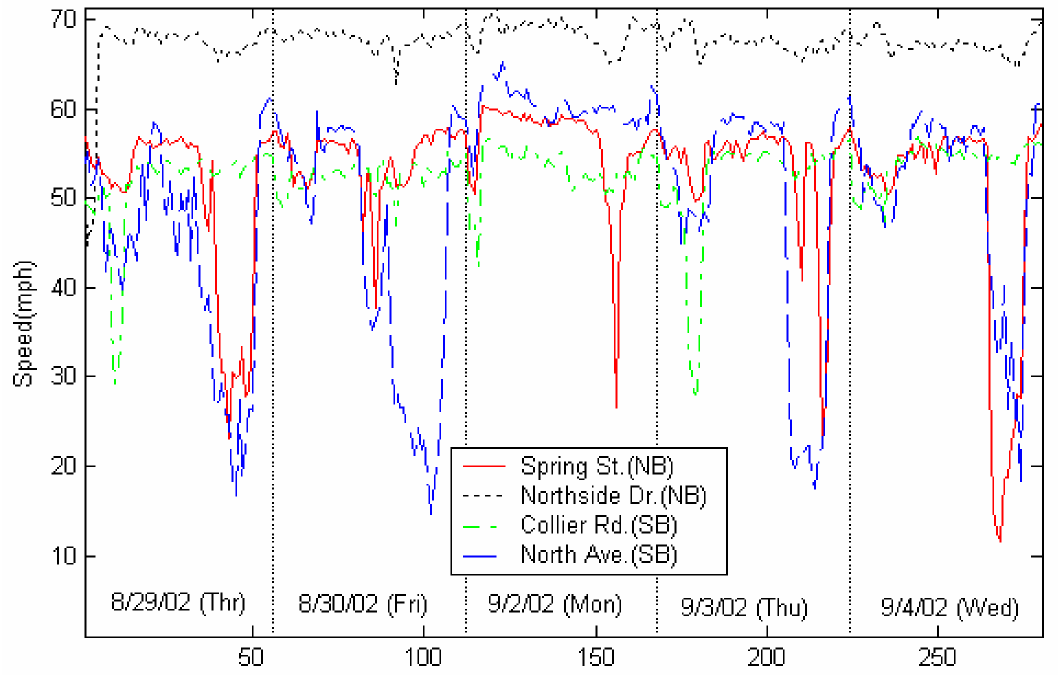
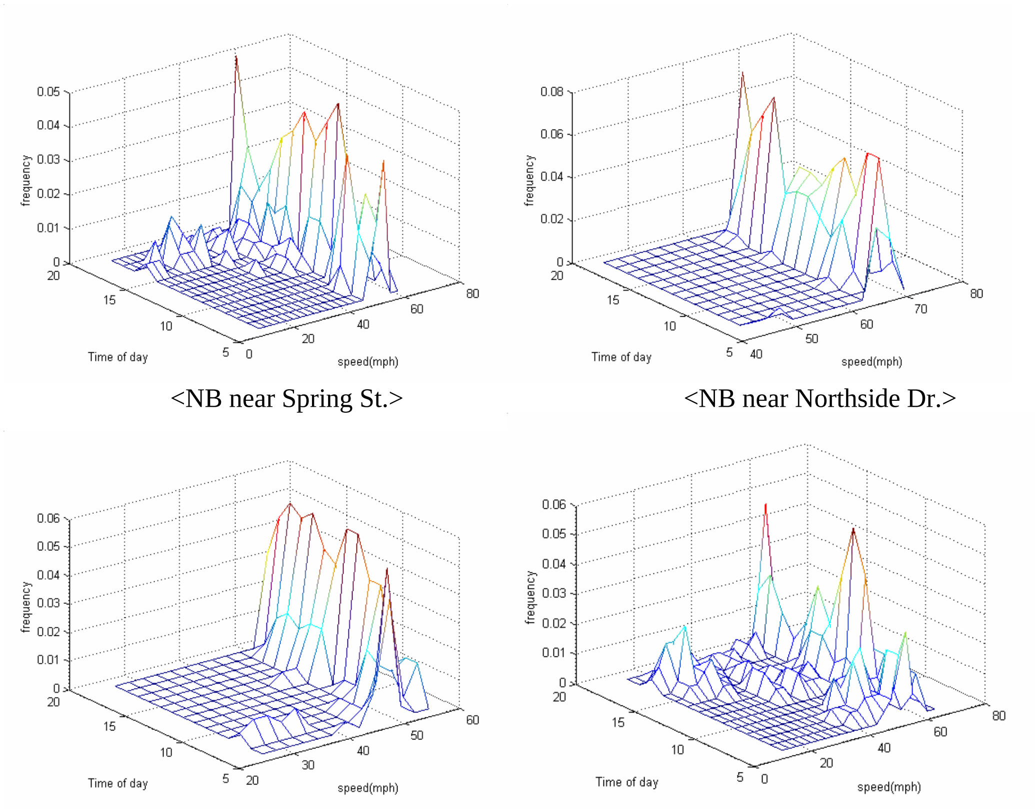


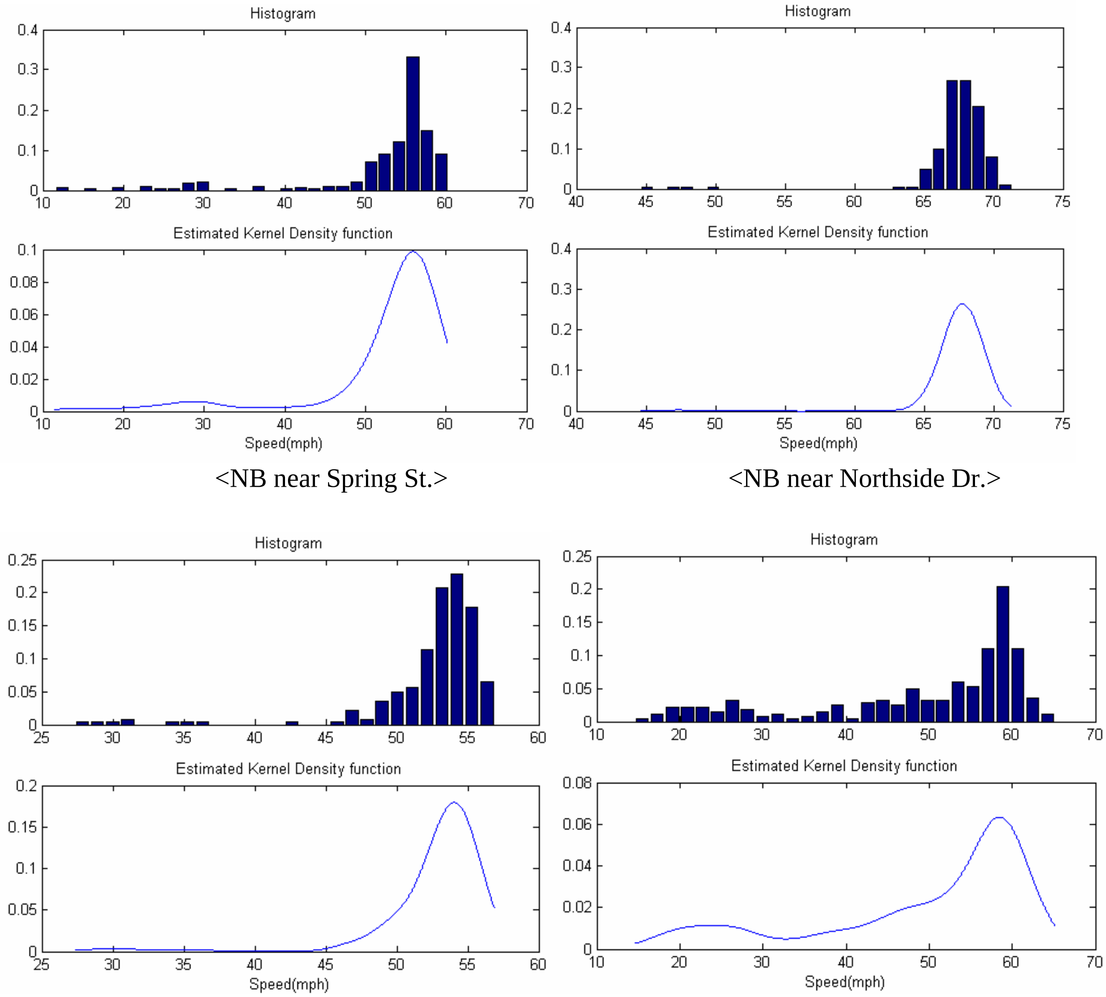
FIGURE 2 Speed profiles during a week



<NB near Spring St.> <NB near Northside Dr.>

<SB near Collier Rd.> <SB near North Ave.>

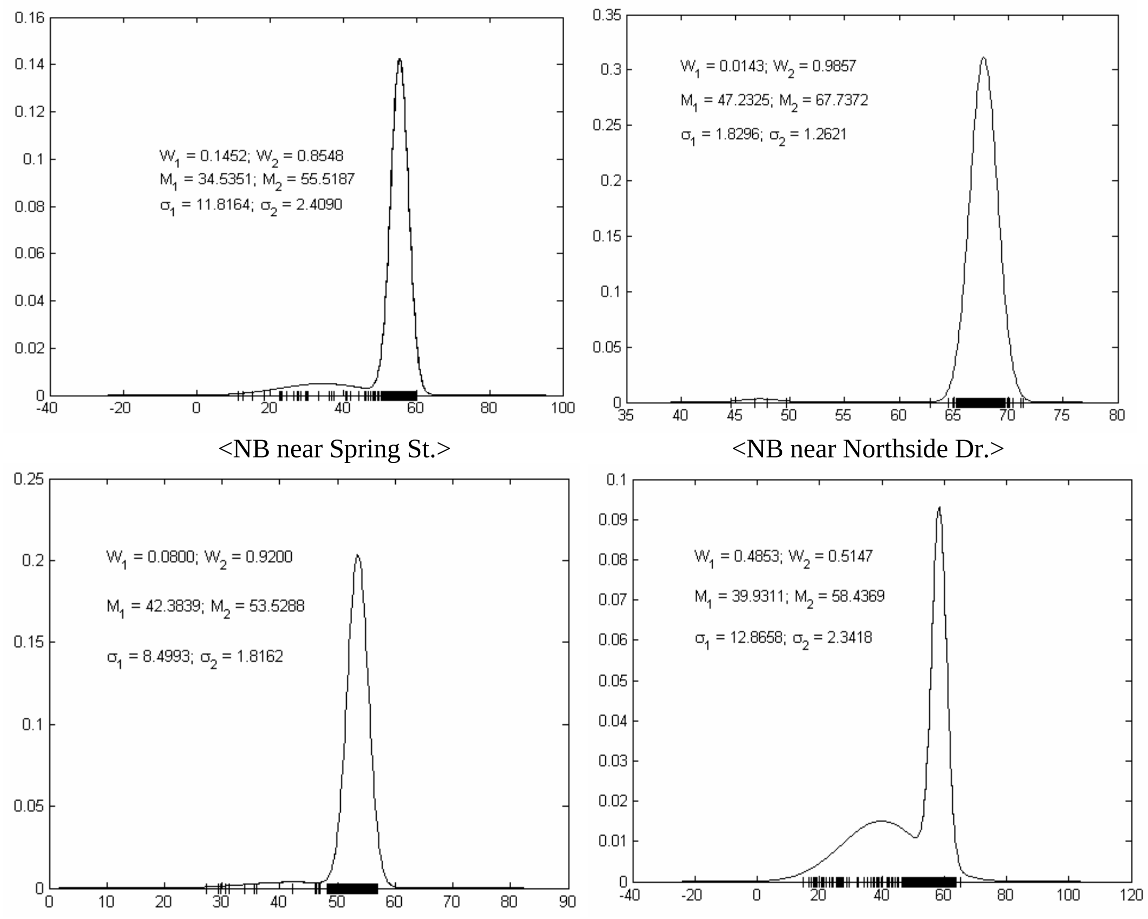
FIGURE 3 Surface plots of speed and time of day



<NB near Spring St.> <NB near Northside Dr.>

<SB near Collier Rd.> <SB near North Ave.>

FIGURE 4 Kernel density estimation of travel speed



<NB near Spring St.> <NB near Northside Dr.>

<SB near Collier Rd.> <SB near North Ave.>

FIGURE 5 Estimated Gaussian mixtures of travel speed

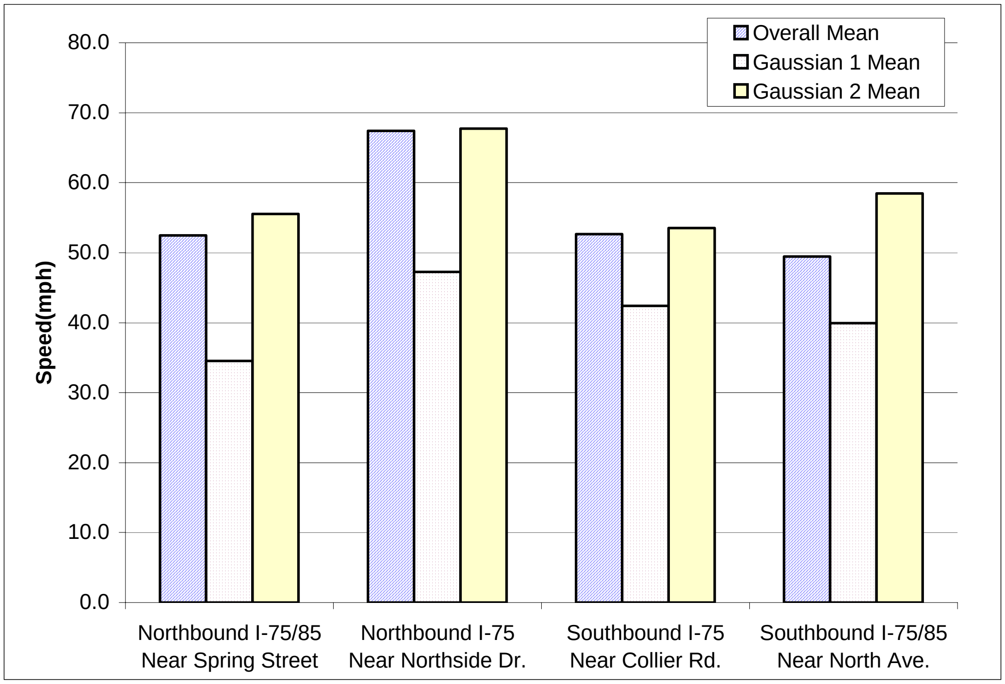


FIGURE 6 Comparison of means from Gaussian mixture model

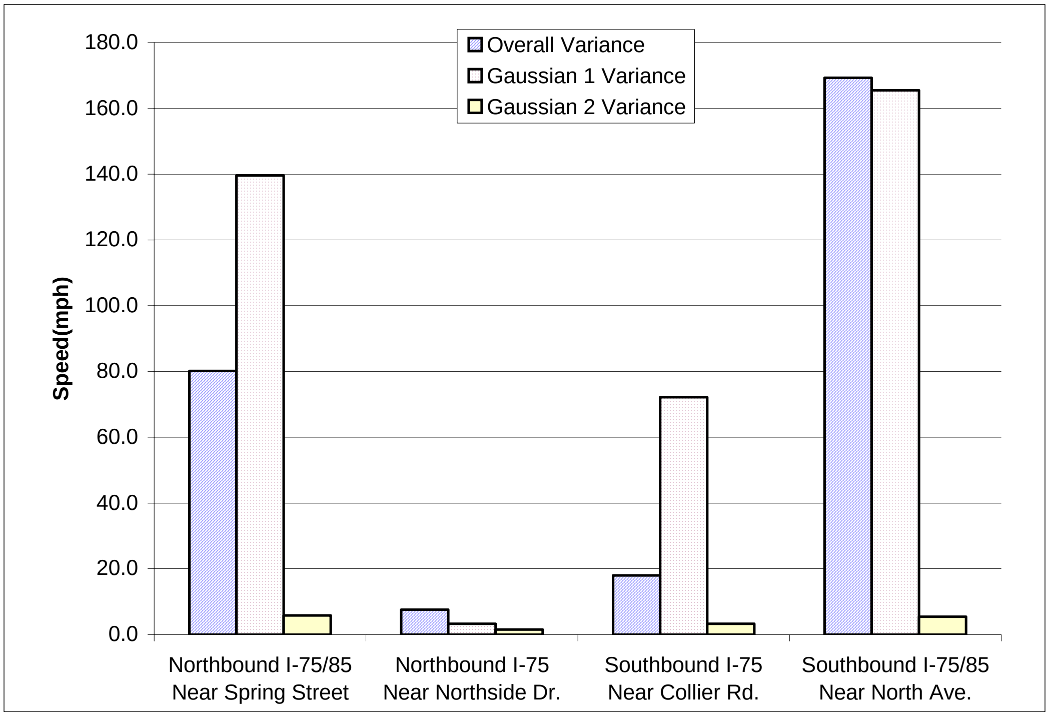


FIGURE 7 Comparison of variances from Gaussian mixture model

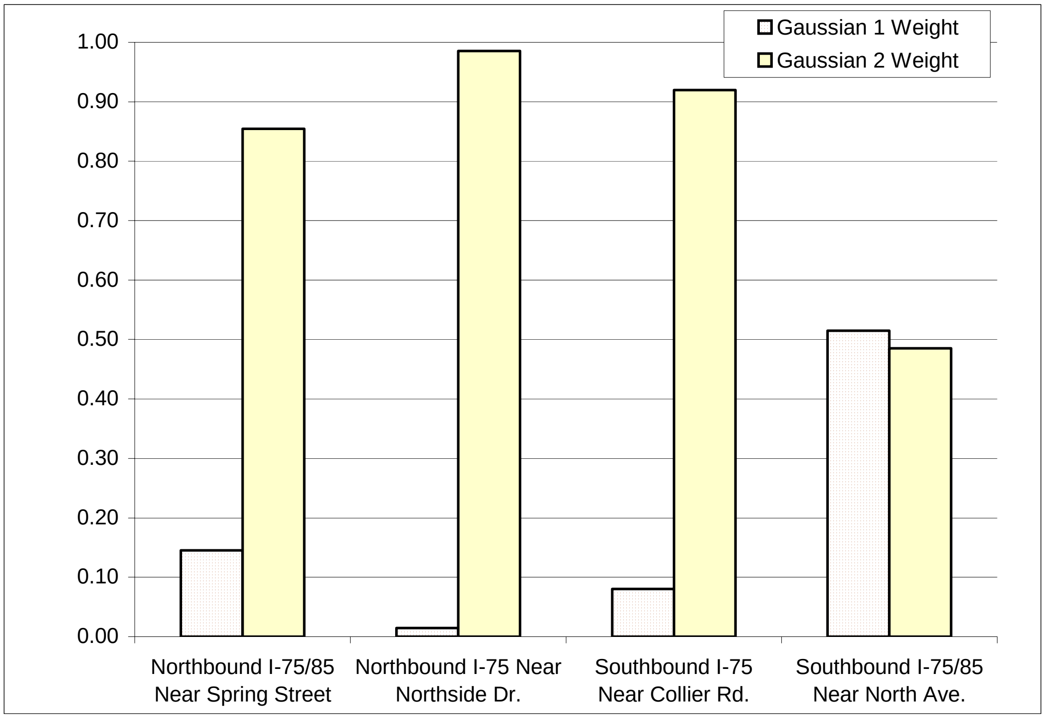


FIGURE 8 Comparison of composition weights of speed distributions