applied sciences

MDPI

Article

Bayesian Mixture Model to Estimate Freeway Travel Time under Low-Frequency Probe Data

Hyungjoo Kim [1,*] and Lanhang Ye [2]

[1] Intelligent Transportation System Laboratory, Advanced Institute of Convergence Technology, Suwon-si 16229, Korea
[2] College of Engineering, Zhejiang Normal University, 688 Yingbin Road, Jinhua 321004, China; ylhlxyz@gmail.com
* Correspondence: hyungjoo@snu.ac.kr

Abstract: This study develops a novel estimation method under low-frequency probe data using the Bayesian approach. Given the challenges in estimating travel time under low-frequency probe data and prior distribution of the parameters in a traditional Bayesian approach, the proposed algorithm adopts a historical data-based data-driven method according to the characteristics of travel time regularity. Due to the variability of travel times during peak periods, this paper adopts a mixture distribution of travel times in the Bayesian approach rather than traditional single distribution. The Gibbs sampling method with a burn-in period is used to generate a series of sampling sequences from an unknown joint posterior distribution for estimating the posterior distribution of the parameters. The proposed algorithm is tested using traffic data collected from the Korean freeway section from Giheung IC to Dongtan IC. Both MAPE and RMSE of the estimation results show that the proposed method has the smallest deviation from the ground truth travel time compared to the simple mean and moving average methods. Moreover, the proposed Bayesian estimation yields the smallest standard deviation of MAPE for all test days. The credible intervals for estimated travel times show that the proposed method provides good accuracy in estimating travel time reliability.

Keywords: Bayesian mixture estimation; low-frequency probe data; data-driven method; individual travel data; credible interval

Citation: Kim, H.; Ye, L. Bayesian Mixture Model to Estimate Freeway Travel Time under Low-Frequency Probe Data. *Appl. Sci.* **2022**, *12*, 6483. https://doi.org/10.3390/app12136483

Academic Editors: Javier Alonso Ruiz and Angel Llamazares

Received: 6 June 2022
Accepted: 22 June 2022
Published: 26 June 2022

1. Introduction

Travel time is one of the most important metrics to measure the operational efficiency of a transportation network; it is of interest to traffic operators and travelers alike. From an operator's perspective, travel time information is used for better management and control of the traffic system to ease congestion. From a traveler's perspective, travel time information can provide them with better route choice and departure time decisions.

Nowadays, most traffic monitoring systems are based on point detectors (e.g., loop detectors) installed along the roadway. Although these systems collect traffic counts and occupancy rates that can be used to estimate point speeds and segment travel times, such estimations are prone to major errors [1–3]. The recent proliferation of in-vehicle electronics (e.g., smartphones with GPS units and electronic toll collection transponders) provides new possible solutions for estimating travel time. These devices on the vehicles are used to measure actual travel times between distant locations; the actual travel time can be calculated after the vehicle has passed a road segment [4–9].

Although estimation-based probe vehicles provide more reliable information to capture actual travel time, there are still some issues that remain to be resolved. For instance, there are issues with the privacy and low penetration rate of on-board devices [10–15]. If a sufficient number of samples cannot be obtained to estimate travel times within a time interval, biased travel times can be estimated, with higher variance [12,13,16].

Given this situation, several different statistical methods have been proposed over the past few years. The regression model is proposed, which makes use of correlations between links to generalize low-frequency probe vehicle data and captures the underlying factors behind spatial and temporal variations in velocity [17]. The Fuzzy logic is also proposed to assign vehicle trajectories to a certain degree of membership. In order to estimate the mean travel time for the entire population, different traffic conditions are matched according to the degree of membership [18–20]. Other previous studies have focused on the effect of small samples from probes on the estimation of mean travel time [13,21–23].

These studies show that simple travel time averages based on low-frequency probes cannot approach the overall mean travel time due to correlations between samples. It must be said that it is still common to obtain a sample of probe vehicles that is small and not representative of the entire population.

As traffic becomes non-stationary, the variability in estimated travel times indicates that the uncertainty and changing characteristics of travel times tend to increase during peak periods [24–26]. Since the distribution of travel time with the variability during peak periods does not follow a single model, mean travel time using point estimation is not meaningful as a statistical inference about the estimator of travel time. In order to be meaningful as a statistical inference, interval estimation of travel time can be used as an alternative to overcome the limitations of the point estimation. However, there are few previous studies on interval estimation of travel time [16,27–29], and it is still difficult to accurately estimate travel time using interval estimation.

The objective of this study is to overcome the aforementioned problems in estimating travel time under low-frequency probe data during peak periods. To this end, this paper develops a novel travel time estimation model using a Bayesian approach. The rest of this article is organized as follows: The next section provides background information on the Bayesian approach and introduces the framework of the novel estimation model. This is followed by a description of the test data for the case study and a comparison of results using different estimation methods. The last section provides the conclusions of the paper and discusses their implications.

2. Methodology

Since the prior distribution of travel time is difficult to quantify, the implementation of traditional Bayesian methods is often challenging in terms of travel time estimation during the time update process. Therefore, this paper proposes a data-driven approach based on historical data to solve this problem. In this section, the novel estimation method in the Bayesian approach is presented as follows:

2.1. Definition of the Bayesian Approach

The Bayesian approach is a statistical inference method in which Bayes' theorem is used to update the probability of a hypothesis as more evidence or information becomes available. The Bayesian approach treats a parameter as a random quantity and is based on the distribution of the parameter conditional on observed data, as provided below:

$$p(\theta|y) \propto L(y|\theta) \cdot \pi(\theta) \quad (1)$$

where $p(\theta|y)$ is the posterior distribution based on given measured data y, $L(y|\theta)$ is the likelihood based on given parameter θ, and $\pi(\theta)$ is the prior distribution before updating with observed data y.

The methodology in this study attempts to estimate accurate travel time under low sampling rates of probe data using the Bayesian approach. According to the characteristics of travel time regularity [30,31], the prior distribution of the parameters can be estimated using a data-driven method. The distribution of travel times during peak periods has a mixture shape of distribution that can be classified into a group of vehicles that have rapidly traveled along the road segment and a group of vehles that have traveled slowly due to the influence of irregular traffic oscillation [32,33]. Since the distribution of travel

time has the mixture shape of distribution, this study adopts a Gaussian mixture model with two components in the Bayesian approach. The Gaussian mixture model has the advantage of expressing the centroid and forming clusters properly even if the variance is not constant. The details of the model with each step included are as follows.

2.2. Likelihood and Model Estimation Method

Estimation of mixture models can be classified into two methods, such as Expectation-Maximization (EM) and Bayesian methods. The EM algorithm has been widely used to estimate the mixture shape of distribution based on the maximum likelihood method. However, this method may lead to a local maximum and many starting points are needed to find a global maximum. Furthermore, since the maximum likelihood method is based on asymptotic theory, the sample size must be large [34,35]. This means that estimating a mixture model is difficult under low sampling rates of probes using the EM algorithm. In this study, a Bayesian mixture model is used as it provides richer inferences than the maximum likelihood method; also, this approach can address parametric uncertainty and travel time properties in estimating travel time. This study employs a mixture of two Gaussian distributions since mixture models are a flexible family of models and have been used to model large heterogeneous populations. The two mixtures of Gaussian distribution with individual travel data, $y = \{y_1, y_2 \cdots y_n\}$ can be expressed as follows:

$$p(y_i|\theta) = \sum_{k=1}^{2} w_k \, N\left(y_i|\mu_k, \sigma_k^2\right) \tag{2}$$

where K is the number of components and w_K is the proportion of component K ($0 \le w_K \le 1$, $\sum_{k=1}^{2} w_K = 1$). $N\left(y_i|\mu_k, \sigma_k^2\right)$ is the normal distribution with mean μ_k and variance σ_k^2. θ is the vector of all parameters, $\theta = \{(w_1, w_2), (\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2)\}$.

The likelihood function of the mixture distribution with latent variable Z for classification can be expressed as follows. Latent variables are variables that are not directly observed but inferred from other observed variables. Thus, the latent variable Z follows the multinomial distribution and the likelihood function is expressed with the latent variable.

$$(y, Z|\theta) = \prod_{i=1}^{n} p(y_i|Z_i, \theta) p(Z_i|\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} [w_k p\left(y_i|\mu_k, \sigma_k^2\right)]^{Z_{i,k}} \tag{3}$$

where $Z_{i,k}$ is the ith element of the kth component and, if y_i follows N (μ_k, σ_k^2), $Z_{i,k}$ is 1; if not, $Z_{i,k}$ is 0.

The Bayesian mixture model is applied when the number of components is assumed to be known. According to the influence of traffic oscillation, this paper used the Bayesian mixture model with two components to estimate the travel time distribution. To classify the individual travel data of the kth component, the conditional probability function can be expressed as follows:

$$p\left(Z_i = k|w^{(r-1)}, \mu^{(r-1)}, \sigma^{2(r-1)}, c\right) \propto w^{(r-1)} \cdot p\left(y_i \mid \mu^{(r-1)}, \sigma^{2(r-1)}\right), \quad \text{for } k = 1, 2, \cdots, K \tag{4}$$

where $w^{(r-1)}, \mu^{(r-1)}, \sigma^{2(r-1)}$ are the parameters of the Gaussian mixture model at $r-1$ iterations and y_i is the individual travel data for classification.

2.3. Prior Distribution Based on Data-Driven Method

In Bayesian methods, the prior distribution of the parameters must be specified. The prior distributions in this study assume conjugate form since the conjugate prior distribution, which is in the same probability distribution family as the posterior distribution, offers a closed-form posterior distribution [35]. In order to estimate prior distributions on the parameters with two components, this paper used the k-means clustering method. The

k-means clustering is popular for cluster analysis in data mining and has a low computational cost.

$$w_k \sim \text{Dirichlet}(e_1, \cdots, e_K) \quad (5)$$

$$\tau_k \sim \text{Gamma}(\alpha, \beta) \quad (6)$$

$$\mu_k|\tau_k \sim N(\mu_0, n_0\tau) \quad (7)$$

where w_k, τ_k, and μ_k are the parameters of the Gaussian mixture distribution. Dirichlet is the conjugate prior for weight distribution and $e_1, \cdots, e_K$ are the hyper-parameters of the prior distribution. Gamma is the conjugate prior for precision distribution and α, β are the hyper-parameters of the prior distribution. N is the conjugate prior for mean distribution and $\mu_0, n_0\tau$ are the hyper-parameters of the prior distribution.

The hyper-parameters of w_k, τ_k, and μ_k in this study are estimated by the data-driven method based on historical data according to the characteristics of travel time regularity [30,31]. The data-driven method is a prediction algorithm used to compare the distance between current and historical neighbors to find neighbors that are closest to the current states. The method assumes that traffic conditions similar to the current one exist in the past and usually estimates state values using a large amount of historical traffic data to select the candidates through the Euclidean distance with the current data sequence. It assumes current and historical travel time sequences, which are denoted by tail time; the distance between two sequences is calculated using the Euclidean distance, which represents the dissimilarity measure between these two sequences. The data-driven method has been used in numerous studies for predicting short-term traffic conditions [36–40]. For the hyper-parameters of prior distribution in this study, the optimal number of hyper-parameters of prior distribution was calculated through the trial-and-error method. The number of hyper-parameters of prior distribution with the minimum error was selected, and 10 was selected, and to estimate prior distributions on the parameters with two components, k = 2 is chosen in k-means clustering.

The data sequence length of the data-driven method is chosen to be 3 periods (15 min), which entails travel time patterns in short-term traffic flow. To increase accuracy and to decrease computational cost, the k-candidates in the data-driven method are selected while searching past data in the same time slot.

$$\text{Min. Dist} = \sqrt{\sum_{i=1}^{n}\left(T_i^c - T_i^h\right)^2}, \text{ “for all samples”} \quad (8)$$

where T_i^c is the ith current travel time sequence, T_i^h is the ith historical travel time sequence, n is the length of travel time sequence, and Dist is a dissimilarity measure at time t between the two sequences, obtained using the Euclidean distance.

2.4. Posterior Distribution Based on Gibbs Sampling Method

Posterior distribution, which is augmented with the component indicator Z, is estimated by likelihood and prior distribution. The posterior distribution is in the same probability distribution family as the prior probability distribution since this paper used conjugate form of prior distribution; the details of derivations are expressed as follows:

$$p(w|y) \propto L(y|w)\cdot\pi(w) \propto \frac{n!}{\prod_i n_i!}\prod_{i=1}^{K} w_i^{n_i}\cdot\frac{1}{B(e)}\prod_{i=1}^{K} w_i^{e_i-1} = \prod_{i=1}^{K} w_i^{e_i+n_i-1} \quad (9)$$

$$B(e) = \frac{\prod_{i=1}^{k}\Gamma(e_i)}{\Gamma\left(\sum_{i=1}^{k} e_i\right)}, e = (e_1, \cdots, e_K)$$

where Γ is the gamma function as a normalizing constant, e_i is the hyper-parameter of prior distribution, and n_i is the number of observations allocated to component k.

$$p(\tau, \mu|y) \propto L(y|\tau, \mu)\cdot\pi(\mu|\tau)\cdot\pi(\tau) \quad (10)$$

$$\propto \tau^{n/2} \exp\Big(-\frac{\tau}{2}\sum(y_i-\mu)^2\Big)\cdot\tau^{1/2}\exp\Big(-\frac{n_0\tau}{2}(\mu-\mu_0)^2\Big)\cdot\tau^{\alpha-1}e^{-\beta\tau}$$

$$\propto \tau^{\alpha+\frac{n}{2}-1}\exp\Bigg(-\tau\Big(\beta+\frac{1}{2}\sum(y_i-\bar{y})^2\Big)\Bigg)\tau^{1/2}\exp\Big(-\frac{\tau}{2}\Big(n_0(\mu-\mu_0)^2+n(\bar{y}-\mu)^2\Big)\Big)$$

which leads to a gamma posterior for τ:

$$P(\tau|y)\propto\tau^{\alpha+\frac{n}{2}-1}\exp\Bigg(-\tau\Bigg(\beta+\frac{1}{2}\sum(y_i-\bar{y})^2+\frac{nn_0}{2(n+n_0)}(\bar{y}-\mu_0)^2\Bigg)\Bigg)$$

$$\tau|y\ \sim\ \text{Gamma}\Bigg(\alpha+\frac{n}{2},\ \beta+\frac{1}{2}\sum(y_i-\bar{y})^2+\frac{nn_0}{2(n+n_0)}(\bar{y}-\mu_0)^2\Bigg)$$

where $\bar{y}$ is the mean of the observations in component k, μ_0, n_0 are parameters of the prior distribution for the mean, and $\alpha,\ \beta$ are parameters of the prior distribution for precision.

$$p(\mu|y)\ \propto\ L(y|\mu)\cdot\pi(\mu) \tag{11}$$

$$\propto\ \exp(-\frac{\tau}{2}\textstyle\sum_{i=1}^{n}(y_i-\mu)^2)\ \exp(-\frac{n_0\tau}{2}(\mu-\mu_0)^2)$$

$$\propto\ \exp(-\frac{n\tau}{2}(\bar{y}-\mu)^2)\ \exp(-\frac{n_0\tau}{2}(\mu-\mu_0)^2)$$

$$\propto\ \exp(-\frac{n\tau}{2}(\mu^2-2\bar{y}\,\mu)-\frac{n_0\tau}{2}(\mu^2-2\mu_0\mu)^2)$$

$$\propto\ \exp(-\frac{1}{2}(n\tau+n_0\tau)(\mu^2-2\frac{(n\tau\,\bar{y}+n_0\tau\mu_0)}{(n\tau+n_0\tau)}\mu))$$

$$\propto\ \exp(-\frac{1}{2}(n\tau+n_0\tau)(\mu-\frac{(n\tau\,\bar{y}+n_0\tau\mu_0)}{(n\tau+n_0\tau)})^2)$$

$$\mu|y\ \sim\ N(\frac{n\tau}{n\tau+n_0\tau}\bar{y}+\frac{n_0\tau}{n\tau+n_0\tau}\mu_0, n\tau+n_0\tau)$$

where $\bar{y}$ is the mean of the observations in component k, n is the number of observations allocated to component k, and μ_0, n_0 are parameters of the prior distribution for mean. The posterior distributions of the proposed Bayesian mixture model are expressed as in Table 1.

Table 1. Posterior distribution of the proposed Bayesian mixture model.

Parameter	Likelihood	Conjugate Prior	Posterior
w	$y_i\|w\ \sim$ Mulinomial (w)	$w\ \sim$ Dirichelt $(e_1,\cdots,e_K)$	$w\|Z,y\ \sim$ Dirichelt $(e_1+n_1,\ \cdots,e_K+n_K)$
$\tau=\frac{1}{\sigma^2}$	$y_i\|\mu,\tau\ \sim$ N (μ,τ)	$\mu\|\tau\ \sim$ N $(\mu_0,n_0\tau)$ $\tau\ \sim$ Gamma (α,β)	$\tau\|Z,y\ \sim$ Gamma $\Big(\alpha+\frac{n}{2},\ \beta+\frac{1}{2}\sum(y_i-\bar{y})^2+\frac{nn_0}{2(n+n_0)}(\bar{y}-\mu_0)^2\Big)$
μ	$y_i\|\mu,\tau\ \sim$ N (μ,τ) with known τ	$\mu\|\tau\ \sim$ N $(\mu_0,n_0\tau)$	$\mu\|Z,\tau,y\ \sim$ N $\Big(\frac{n\tau}{n\tau+n_0\tau}\bar{y}+\frac{n_0\tau}{n\tau+n_0\tau}\mu_0, n\tau+n_0\tau\Big)$

Although the proposed mixture model is based on a standard distribution, the inference of the posterior distribution of the model parameters is analytically intractable (McLachlan et al., 2000; Gelman, et al., 2003). Fortunately, due to new computational techniques, it has recently become possible to estimate mixture models using the Markov Chain Monte Carlo (MCMC) algorithms. The Gibbs sampling method is a special MCMC algorithm that is widely used to generate sample draw sequences from an unknown joint posterior distribution. During each iteration of the algorithm, samples of each parameter are alternately drawn from the conditional posterior distribution, given the other parameters drawn recently. If the sequence is long enough, it can be used to estimate the joint distribution. This paper used the Gibbs sampling method including the burn-in period to estimate the posterior distributions of model parameters. We first use R_0 to draw the

burn-in period, with iteration value R. We perform R = 15,000 iterations and discarded the first R_0 = 5000 draws as a burn-in period, then estimated the parameters as follows:

$$\widehat{w}_k = \frac{1}{R - R_0} \sum_{r=R_0+1}^{R} {w_k}^{(r)} \tag{12}$$

$$\widehat{\mu}_k = \frac{1}{R - R_0} \sum_{r=R_0+1}^{R} {\mu_k}^{(r)} \tag{13}$$

$$\widehat{\sigma_k^2} = \frac{1}{R - R_0} \sum_{r=R_0+1}^{R} \sigma_k^{2(r)} \tag{14}$$

This paper sets an initial allocation of $Z^{(0)}$, with initial values for $\mu_k^{(0)}$ and $\tau_k^{(0)}$ to classify the individual travel data. For the initial allocation $Z^{(0)}$, we assumed that there is an equal prior for the prior weights assumed and for the initial $\mu_k^{(0)}$ and $\tau_k^{(0)}$, the parameters of each clustering step are calculated at the first iteration step. The framework of the proposed Bayesian mixture estimation is summarized as follows:

This paper uses initial allocation $Z^{(0)}$, with initial values for $\mu_k^{(0)}$ and $\tau_k^{(0)} = \frac{1}{\sigma_k^{2\,(0)}}$, and repeat the following steps for $r = 1, \cdots, R_0, \cdots, R_0 + R$.

Step 1: Classification, $Z^{(r)}$ with conditional on knowing $(w^{(r-1)}, \mu^{(r-1)}, \sigma^{2(r-1)})$

Classify each observation y_i for $i = 1, \cdots, N$, with the following probability rule

$$p(Z_i = k|\ \mu,\ \sigma^2,\ w,\ y_i) \propto \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{(y_i - \mu_k)^2}{2\sigma_k^2} \right\} \cdot w_k$$

Step 2: Parameter estimation, $(w^{(r)}, \mu^{(r)}, \sigma^{2(r)})$

Sample $w^{(r)}$ from posterior distribution of Dirichlet

Sample $\tau_k^{(r)}$ for each component k from posterior distribution of Gamma

Sample $\mu_k^{(r)}$ for each component k from posterior distribution of Normal

Store the values of all parameters as $\theta^{(r)} = \left\{ \left(w_1^{(r)}, \cdots, \mu_k^{(r)} \right), \left(\mu_1^{(r)}, \cdots, \mu_k^{(r)} \right), \left(\tau_1^{(r)}, \cdots, \tau_k^{(r)} \right) \right\}$

Increase r by one, and return to Step 1

Step 3: Discard the first R_0 draws as a burn-in period.

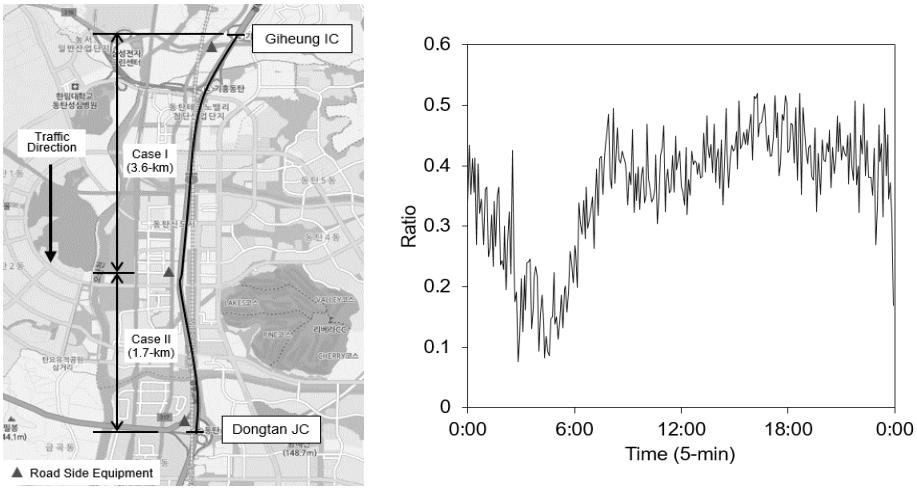
3. Model Evaluation

In this section, an empirical study is performed to evaluate the proposed travel time estimator. The test environment is first introduced, followed by implementation of different estimation methods on the same test dataset. Finally, the test results and discussion are given.

3.1. Descriptions of Site and Data

The study site used for empirical analysis is two segments of the Busan-bound Gyeongbu Expressway connecting Giheung IC to Dongtan IC, as illustrated in Figure 1a. The selected freeway sections typically experience high traffic volume and heavy congestion on weekends since they are main routes to tourist attractions. Therefore, travelers need efficient and accurate travel time estimates when planning trips and changing travel routes, especially during peak periods. The evaluation of travel time estimation on the test site is conducted based on individual probe data from Korean Expressway Corporation. The dataset provided by Korean Expressway Corporation is mainly collected by Korea Expressway's ETC system, Hi-Pass, using two different types of transponder technologies, radio frequency (RF) and infrared ray (IR), through dedicated short-range communication (DSRC) (Kim et al., 2013). The individual probe data at the test site covers two freeway segments with a total length of 5.3-km. The raw probe data provides individual travel times

for each segment and are collected in different time intervals. In this study, the raw data are used at five-minute intervals to estimate travel time and their distribution of travel times. The individual probe data between 6:30 and 13:00 from 1 March 2017 to 28 April 2017 are used for the training dataset, and from 29 April and 5 May 2017 are used for the test dataset because the most congested periods are observed in this time frame. Therefore, the estimation performances using different methods are investigated during the peak period.



(**a**) (**b**)

Figure 1. Study site and assumption of ground truth: (**a**) Gyeongbu Expressway, Giheung IC-Dongtan JC, (**b**) Ratio of traffic volume between loop detectors and ETC.

To evaluate the estimation accuracy of the methodology, this study assumes the travel time estimated on the Gyeongbu expressway to be the ground truth. The Gyeongbu expressway, which relatively includes a high penetration rate of on-board devices in South Korea, is selected to test the estimation algorithm. The loop detector collects a full volume of traffic during the time interval, and it is possible to compare those data with the amount of traffic volume collected by ETC readers. The traffic volume from loop detectors was compared with the penetration rate of ETC between Giheung IC to Dongtan IC, as illustrated in Figure 1b. The traffic volume of passenger cars was used to compare between loop detectors and ETC, and traffic volume from different types of vehicles was excluded. Since the ratio traffic volume between loop detectors and ETC is approximately 50% in the study site, the estimation results obtained by assuming the travel time collected by ETC as ground truth were compared.

3.2. Comparison of Estimation Methods and Performance Indicators

For performance evaluation, several different methods are also applied on the same data set. The methods chosen include the state-of-the-art simple mean method and the moving average method as comparison groups. The simple mean method is the easiest alternative to estimate travel times by collecting real-time data on the segments. This method is generally used by Korean Expressway Corporation to display travel time information on variable message signs. Therefore, the simple mean method is considered a state-of-the-art method for quantifying the trade-off between simplicity and estimation accuracy. The moving average is often used with time-series data to smooth out short-term fluctuations in real-time travel time estimations by creating a series of averages of different subsets of the full data set to analyze the calculation of data points. The threshold depends on the application, and the parameters of the moving average will be set accordingly. In this study, three and five orders of moving average are selected, which entails travel time patterns in short-term traffic flow.

To evaluate different estimation methods, performance criteria are specified using unbiasedness and efficiency to obtain a good estimator. The unbiasedness means that the estimated value of the estimators should be equal to the true value of the variable estimated. This study adopts both the absolute and root mean square estimation errors for measuring the unbiasedness of the different estimators. The Mean Absolute Percentage Error (MAPE) is the average absolute percentage change between the estimated and the true values. Also, the Root Mean Square Error (RMSE) is also the difference between the estimated and the true values as demonstrated as follows:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|T_i^E - T_i^G|}{T_i^G} \tag{15}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(T_i^E - T_i^G\right)^2} \tag{16}$$

where T_i^E is the estimated travel time of the ith time interval, T_i^G is the ground truth of travel time of the ith time interval, and n is the total number of the estimated travel times.

The efficiency is a measure of the quality of an estimator and an efficient estimate of the good estimator is one that has the smallest standard error among all unbiased estimators. This study adopts the standard deviation of MAPE for measuring the efficiency of the different estimators.

3.3. Estimation Results by Penetration Rate of Probe Data

To evaluate the performance of the novel estimation method during peak periods (6:30–13:00), this study compared average absolute estimation errors by different penetration rate of probes from random sampling in Figure 2. Among all the results, the estimation error has a pattern in which it decreases as the penetration rates of probes increases. Although the estimation error gradually decreases as the number of probes increases, moving average methods show constant errors since the moving average is not intensely related to the penetration rate of probe data. The simple mean method produces similar patterns of estimation error with the Bayesian mixture estimation, but the average absolute estimation error is higher than the Bayesian mixture estimation. The Bayesian mixture estimation shows higher accuracy than other models for a 3–13% penetration rate of probe data. However, the Bayesian mixture estimation has limitations in that it cannot be used to perform estimations at under 1–2% penetration rate of probes due to constraints of the model at low sample numbers.

3.4. Estimation Results under Low Frequency Probe Data

For investigation of the deviation between estimation results and ground truth data, estimation results using scatter plot were produced by the four methods on 3% penetration rate of probe data and presented in Figure 3. The figure clearly shows a significant difference during peak periods (6:30–13:00) between estimated travel time by different methods and ground truth travel time. Among all the methods, the estimated travel times from simple mean and moving average and ground truth travel time highly deviate during peak periods. Meanwhile, Bayesian mixture estimation does not result in significant bias. In conclusion, the figure demonstrates that the estimation errors are occurring during peak periods under low sampling rates of probes.

For better evaluation on the performance of the Bayesian mixture approach, the average absolute and root mean square estimation errors produced during peak periods by the four methods are summarized in Table 2. As shown in the table, the proposed Bayesian mixture method produces the smallest estimation error. Among all the methods, the method of simple mean generally provides the worst performance for all test days, with a significant drop in performance during peak periods. Moving average produces slightly lower estimation errors compared to the simple mean method. Besides, the least standard deviations of MAPE are produced by the proposed Bayesian mixture approach compared

to simple mean and moving average methods. In conclusion, the results show that the proposed algorithm outperforms the different estimation methods during peak periods.

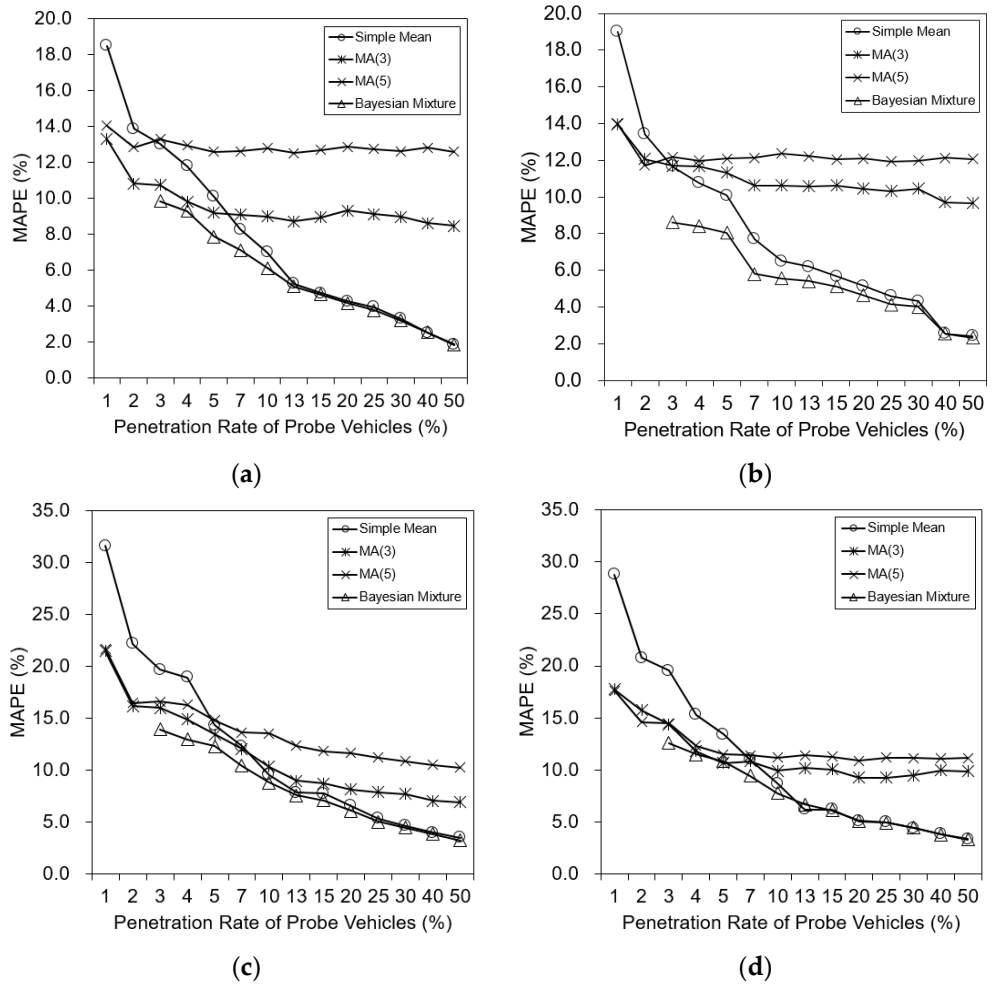


Figure 2. MAPE by different penetration rates of probe data: (**a**) Case I, 29 April 2017, (**b**) Case II, 29 April 2017, (**c**) Case I, 5 May 2017, (**d**) Case II, 5 May 2017.

Table 2. Estimation results by different methods during peak periods.

		Simple Mean	MA(3)	MA(5)	Bayesian Mixture
Case I, 29 April	MAPE (%)	14.1	12.0	14.8	10.3
	RMSE (s)	74.9	61.5	65.6	51.2
	SD of MAPE	13.8	10.8	11.0	9.2
Case II, 29 April	MAPE (%)	11.8	12.8	13.7	8.6
	RMSE (s)	31.9	29.6	32.8	21.2
	SD of MAPE	10.5	9.0	11.3	6.6
Case I, 5 May	MAPE (%)	19.4	19.2	21.5	13.6
	RMSE (s)	130.2	110.1	107.5	87.2
	SD of MAPE	18.4	17.8	18.1	10.9
Case II, 5 May	MAPE (%)	18.8	15.5	16.5	11.2
	RMSE (s)	52.5	41.7	39.8	27.6
	SD of MAPE	13.9	13.2	14.3	7.5

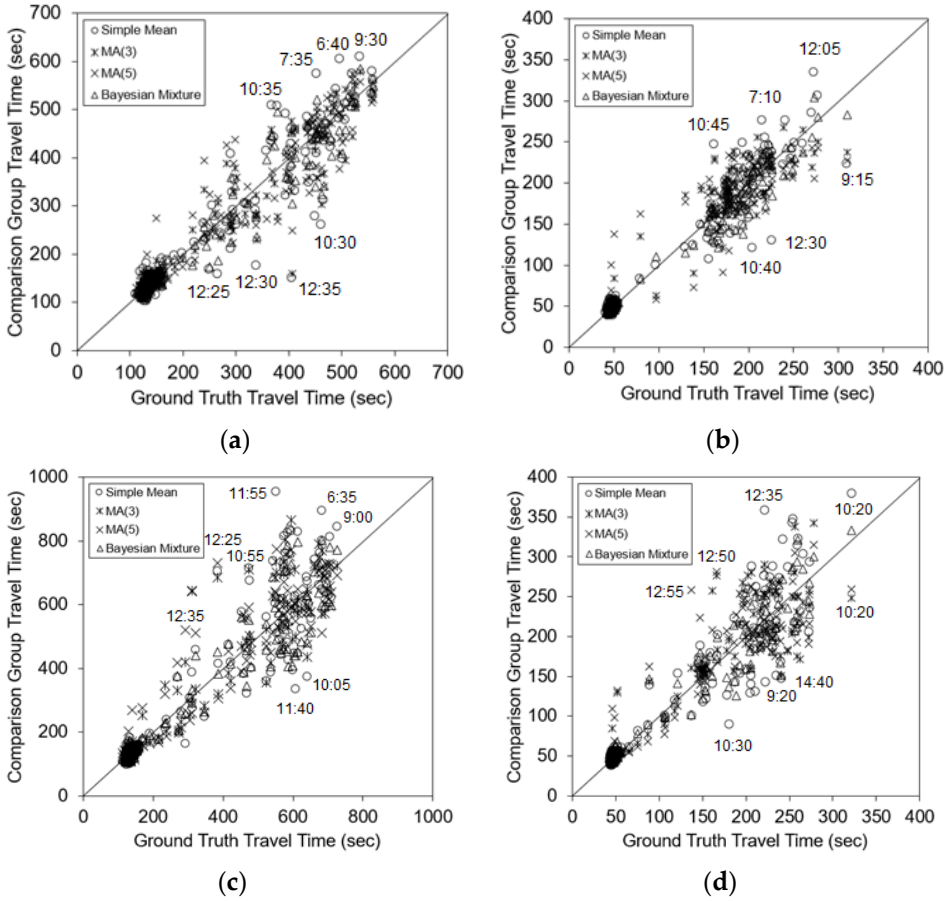


Figure 3. Scatter plot between ground truth and comparison groups: (**a**) Case I, 29 April 2017, (**b**) Case II, 29 April 2017, (**c**) Case I, 5 May 2017, (**d**) Case II, 5 May 2017.

3.5. Credible Intervals under Low Frequency Probe Data

Since the Bayesian mixture not only estimates travel time but can estimate travel time distribution, the credible intervals on 3% penetration rate of probe data are estimated using the Bayesian mixture estimation and presented in Figure 4. The 95% credible intervals of the estimated travel times are calculated using a 2.5% lower boundary and 97.5% upper boundary. Due to the variability of travel times during peak periods, the credible intervals of the estimated travel times are slightly large. However, the 95% credible intervals of the estimated travel times cover most of the temporal variation of the ground truth data, suggesting that the proposed Bayesian mixture provides a good accuracy for estimating travel time reliability.

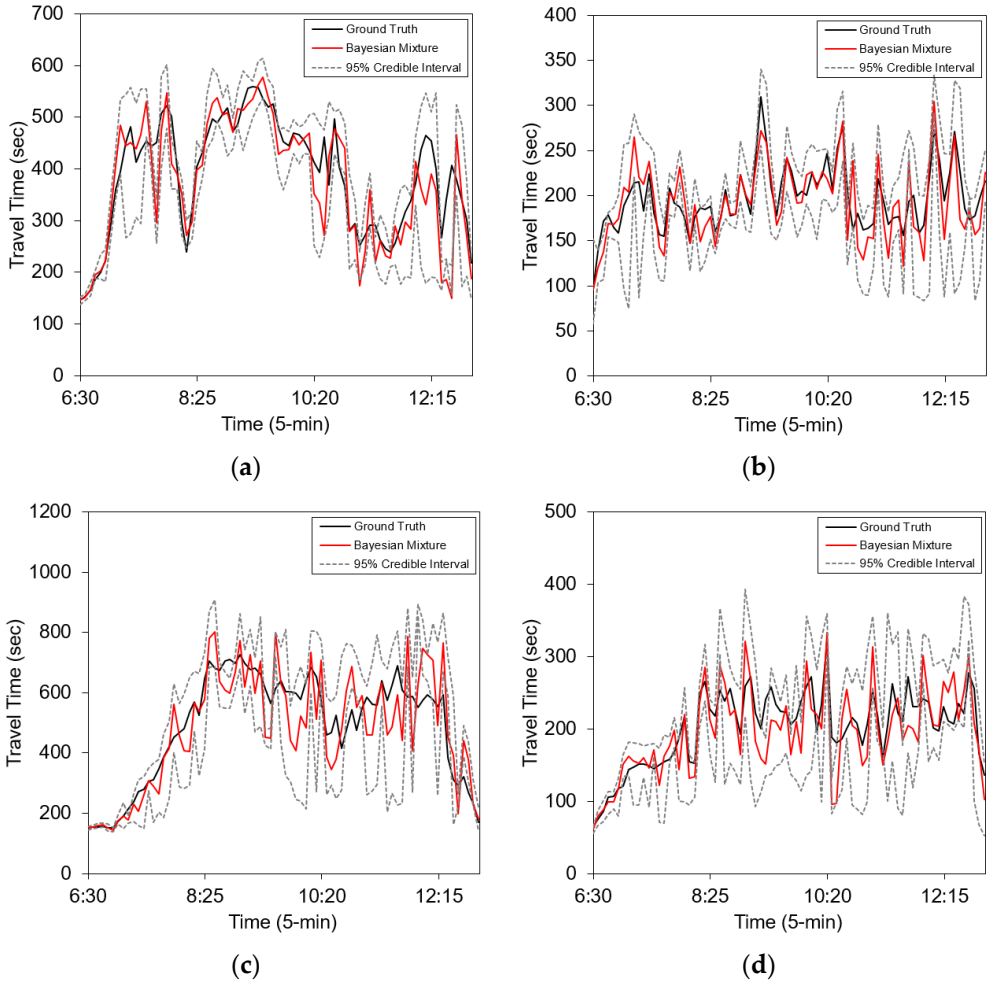


Figure 4. The credible intervals of the estimated travel time on: (**a**) Case I, 29 April 2017, (**b**) Case II, 29 April 2017, (**c**) Case I, 5 May 2017, (**d**) Case II, 5 May 2017.

4. Conclusions

This paper develops a new travel time estimation method under low-frequency probe data based on the Bayesian approach. Due to the variability of travel times during peak periods, this paper adopts a mixture distribution of travel times in the Bayesian approach instead of the conventional single distribution. According to the characteristics of travel time regularity, the proposed estimation method adopts a historical data-based data-driven method to estimate the prior distribution in the Bayesian approach. The Gibbs sampling method, which includes the burn-in period, is used to generate a series of sample draws from an unknown joint posterior distribution and to estimate the posterior distribution of the parameters.

The individual probe data on the selected freeway sections connecting Giheung IC to Dongtan IC are used to study the performance of different estimation methods. To evaluate different estimation methods, the performance criteria are used to specify good estimations using unbiasedness and efficiency. The MAPE and RMSE of estimation results for the unbiasedness show that the proposed method produces the smallest deviation from the ground truth travel times, compared to simple mean and moving average methods. Furthermore, the proposed Bayesian mixture method yields the smallest standard deviation of the MAPE of efficiency compared to the different estimation methods for all test days. Besides, the proposed approach provides good accuracy in estimating travel time reliability according to the credible intervals.

This method can not only estimate the travel time, but also estimate the travel time distribution under the low-frequency sounding data. The method essentially proposed in

the Bayesian approach does not require a specific type of data source and can be applied to low-frequency data problems in other application fields. The proposed approach is also flexible in addressing data estimation problems in other application fields and can potentially yield a relatively high accuracy if sufficient historical data are provided during peak periods. Implementation of the proposed estimator into arterial travel time estimation will be considered in the future. Moreover, future research will consider the use of the other observable factors (e.g., weather conditions) to accurately estimate prior distributions of parameters in data-driven methods.

Author Contributions: Conceptualization, H.K. and L.Y.; methodology, H.K. and L.Y.; software, H.K.; validation, L.Y.; formal analysis, H.K. and L.Y.; writing—original draft preparation, H.K.; writing—review and editing, L.Y.; funding acquisition, H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1C1C1003296).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kim, H.; Kim, Y.; Jang, K. Systematic Relation of Estimated Travel Speed and Actual Travel Speed. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2780–2789. [CrossRef]
2. Kim, H.; Kim, S.; Park, S.; Jang, K. Assessment of Travel Time Estimates based on Different Vehicle Speed Data: Spot Speed vs. Sampled Journey Speed in South Korean expressways. In Proceedings of the 10th International Conference of Eastern Asia Society for Transportation Studies, Taipei, Taiwan, 9–12 September 2013.
3. Li, R.; Rose, G.; Sarvi, M. Evaluation of Speed-Based Travel Time Estimation Models. *J. Transp. Eng.* **2006**, *132*, 540–547. [CrossRef]
4. Haseman, R.J.; Wasson, J.S.; Bullock, D.M. Real-Time Measurement of Travel Time Delay in Work Zones and Evaluation Metrics Using Bluetooth Probe Tracking. *Transp. Res. Rec.* **2010**, *2169*, 40–53. [CrossRef]
5. Gao, S.; Chabini, I. Optimal routing policy problems in stochastic time-dependent networks. *Transp. Res. Part B Methodol.* **2006**, *40*, 93–122. [CrossRef]
6. Puckett, D.D.; Vickich, M.J. *Bluetooth-Based Travel Time/Speed Measuring Systems Development*; Project #09-00-17; University Transportation Center for Mobility (UTCM): College Station, TX, USA, 2010.
7. Haghani, A.; Hamedi, M.; Sadabadi, K.F.; Young, S.; Tarnoff, P. Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors. *Transp. Res. Rec. J. Transp. Res. Board* **2010**, *2160*, 60–68. [CrossRef]
8. Carrese, S.; Cipriani, E.; Crisalli, U.; Gemma, A.; Mannini, L. Bluetooth Traffic Data for Urban Travel Time Forecast. *Transp. Res. Procedia* **2021**, *52*, 236–243. [CrossRef]
9. Pu, Z.; Cui, Z.; Tang, J.; Wang, S.; Wang, Y. Multi-Modal traffic speed monitoring: A real-time system based on passive Wi-Fi and Bluetooth sensing technology. *IEEE Internet Things J.* **2021**. [CrossRef]
10. Rose, G. Mobile Phones as Traffic Probes: Practices, Prospects and Issues. *Transp. Rev.* **2007**, *26*, 275–291. [CrossRef]
11. Srinivasan, K.K.; Jovanis, P.P. Determination of Number of Probe Vehicles Required for Reliable Travel Time Measurement in Urban Network. *Transp. Res. Rec.* **2007**, *1537*, 15–22. [CrossRef]
12. Hellinga, B.R.; Fu, L. Reducing bias in probe-based arterial link travel time estimates. *Transp. Res. Part C Emerg. Technol.* **2002**, *10*, 257–273. [CrossRef]
13. Sen, A.; Thakuriah, P.; Zhu, X.-Q.; Karr, A. Frequency of Probe Reports and Variance of Travel Time Estimates. *J. Transp. Eng.* **1997**, *123*, 290–297. [CrossRef]
14. Cheu, R.L.; Xie, C.; Lee, D.-H. Probe Vehicle Population and Sample Size for Arterial Speed Estimation. *Comput. Civ. Infrastruct. Eng.* **2002**, *17*, 53–60. [CrossRef]
15. Gheorghiu, R.; Iordache, V.; Cormoș, A. Analysis of the Possibility to Detect Road Vehicles via Bluetooth Technology. *Sensors* **2021**, *21*, 7281. [CrossRef] [PubMed]
16. Shi, C.; Chen, B.Y.; Li, Q. Estimation of Travel Time Distributions in Urban Road Networks Using Low-Frequency Floating Car Data. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 253. [CrossRef]
17. Jenelius, E.; Koutsopoulos, H.N. Travel Time Estimation for Urban Road Networks Using Low Frequency Probe Vehicle Data. *Transp. Res. Part B Methodol.* **2013**, *53*, 64–81. [CrossRef]
18. Li, Y.; Mike, M. Link Travel Time Estimation Using Single GPS Equipped Probe Vehicle. In Proceedings of the IEEE 5th International Conference on Intelligent Transportation Systems, Singapore, 3–6 September 2002; pp. 932–937.
19. Lee, S.; Viswanathan, M.; Yang, Y. A Hybrid Soft Computing Approach to Link Travel Speed Estimation. In Proceedings of the Fuzzy Systems and Knowledge Discovery, Third International Conference, FSKD 2006, Xi'an, China, 24–28 September 2006; pp. 794–802.

20. Zhang, X.F.; Li, R.M.; Liu, M.; Shi, Q.X. Evaluating Travel Time Reliability Based on Fuzzy Logic. *Appl. Mech. Mater.* **2011**, *97–98*, 952–955. [CrossRef]
21. Hellinga, B.; Fu, L. Assessing Expected Accuracy of Probe Vehicle Travel Time Reports. *J. Transp. Eng.* **1998**, *125*, 524–530. [CrossRef]
22. Oh, J.-S.; Jayakrishnan, R. Emergence of Private Advanced Traveler Information System Providers and Their Effect on Traffic Network Performance. *Transp. Res. Rec. J. Transp. Res. Board* **2002**, *1783*, 167–177. [CrossRef]
23. Zhou, X.; Yang, Z.; Zhang, W.; Tian, X.; Bing, Q. Urban Link Travel Time Estimation Based on Low Frequency Probe Vehicle Data. *Discret. Dyn. Nat. Soc.* **2016**, *2016*, 7348705. [CrossRef]
24. Mahmassani, H.S.; Hou, T.; Dong, J. Characterizing Travel Time Variability in Vehicular Traffic Networks: Deriving a Robust Relation for Reliability Analysis. *Transp. Res. Rec.* **2012**, *2315*, 141–152. [CrossRef]
25. Bauer, D.; Tulic, M.; Scherrer, W. Modelling travel time uncertainty in urban networks based on floating taxi data. *Eur. Transp. Res. Rev.* **2019**, *11*, 46. [CrossRef]
26. Spreafico, C.; Russo, D. Exploiting the Scientific Literature for Performing Life Cycle Assessment about Transportation. *Sustainability* **2020**, *12*, 7548. [CrossRef]
27. van Hinsbergen, C.; van Lint, J.; van Zuylen, H. Bayesian committee of neural networks to predict travel times with confidence intervals. *Transp. Res. Part C Emerg. Technol.* **2009**, *17*, 498–509. [CrossRef]
28. Park, B.; Zhang, Y.; Lord, D. Bayesian mixture modeling approach to account for heterogeneity in speed data. *Transp. Res. Part B* **2010**, *44*, 662–673. [CrossRef]
29. Jintanakul, K.; Chu, L.; Jayakrishnan, R. Bayesian Mixture Model for Estimating Freeway Travel Time Distributions from Small Probe Samples from Multiple Days. *Transp. Res. Rec. J. Transp. Res. Board* **2009**. [CrossRef]
30. Chen, H.; Rakha, H.; McGhee, C. Dynamic travel time prediction using pattern recognition. In Proceedings of the 20th World Congress on Intelligent Transportation Systems, Tokyo, Japan, 14–18 October 2013.
31. Kumar, B.; Vanajakshi, L.; Subramanian, S. Pattern-Based Bus Travel Time Prediction under Heterogeneous Traffic Conditions. In Proceedings of the 93rd Transportation Research Board Annual Meeting, Washington, DC, USA, 12–16 January 2014.
32. Laval, J.A.; Chen, D.; Ben Amer, K.; Guin, A.; Ahn, S. Evolution of oscillations in congested traffic: Improved estimation method and additional empirical evidences. *Transp. Res. Rec.* **2009**, *2124*, 194–202. [CrossRef]
33. Kim, H.; Jang, K. Characteristics of Travel Time Variability in Congested Traffic. In Proceedings of the 23rd ITS World Congress, Melbourne, Australia, 10–14 October 2016.
34. McLachlan, G.; Peel, D. *Finite Mixture Models*; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2000.
35. Frühwirth-Schnatter, S. *Finite Mixture and Markov Switching Models*; Springer Series in Statistics; Springer: New York, NY, USA, 2006.
36. Smith, B.L.; Williams, B.M.; Oswald, R.K. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transp. Res. Part C Emerg. Technol.* **2002**, *10*, 303–321. [CrossRef]
37. Clark, S. Traffic Prediction Using Multivariate Nonparametric Regression. *J. Transp. Eng.* **2003**, *129*, 161–168. [CrossRef]
38. Qiao, W.; Haghani, A.; Hamedi, M. A Nonparametric Model for Short-Term Travel Time Prediction Using Bluetooth Data. *J. Intell. Transp. Syst.* **2012**, *17*, 165–175. [CrossRef]
39. Tak, S.; Kim, S.; Jang, K.; Yeo, H. Real-Time Travel Time Prediction Using Multi-level k-Nearest Neighbor Algorithm and Data Fusion Method. In Proceedings of the Computing in Civil and Building Engineering, American Society of Civil Engineers, Orlando, FL, USA, 23–25 June 2014.
40. Zhong, J.; Ling, S. Key Factors of k-Nearest Neighbor Nonparametric Regression in Short-Time Traffic Flow Forecasting. In *Proceedings of the 21st International Conference on Industrial Engineering and Engineering Management 2014*; Atlantis Press: Amsterdam, The Netherlands, 2015.