Hindawi
Journal of Sensors
Volume 2017, Article ID 1321237, 8 pages
https://doi.org/10.1155/2017/1321237

Research Article

Operating Time Division for a Bus Route Based on the Recovery of GPS Data

Jian Wang and Yang Cao

School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150090, China

Correspondence should be addressed to Yang Cao; caoyang_202@163.com

Received 24 April 2017; Revised 22 June 2017; Accepted 11 July 2017; Published 14 August 2017

Academic Editor: Xiaolei Ma

Bus travel time is an important source of data for time of day partition of the bus route. However, in practice, a bus driver may deliberately speed up or slow down on route so as to follow the predetermined timetable. The raw GPS data collected by the GPS device equipped on the bus, as a result, cannot reflect its real operating conditions. To address this concern, this study first develops a method to identify whether there is deliberate speed-up or slow-down movement of a bus. Building upon the relationships between the intersection delay, link travel time, and traffic flow, a recovery method is established for calculating the real bus travel time. Using the dwell time at each stop and the recovered travel time between each of them as the division indexes, a sequential clustering-based time of day partition method is proposed. The effectiveness of the developed method is demonstrated using the data of bus route 63 in Harbin, China. Results show that the partition method can help bus enterprises to design reasonable time of day intervals and significantly improve their level of service.

1. Introduction

A well-designed bus schedule scheme is important for increasing bus transit ridership [1]. Bus passenger demand differs greatly at different time intervals during the everyday operation. Before the overall design of a bus schedule scheme, the operating time of a bus route should be divided into multiple time intervals for which different schedule schemes should be made. This greatly helps formulate precise operating and dispatching schemes for buses and reduce the operational costs of a bus transit enterprise.

In recent years, buses in a number of large cities in China have been equipped with GPS devices [2–5]. Bus enterprises can now directly retrieve the bus travel time between any two stops from this database. However, given the predetermined timetables, the travel time may not reflect the actual performance of a bus. When faced with traffic jams, a bus driver may deliberately accelerate if the bus is to arrive at a downstream stop no later than the scheduled time. Although it may manage to arrive on time, the bus typically undergoes frequent acceleration and deceleration enroute which not only reduces the comfort of passengers but also increases the probability of traffic accidents. In contrast, the bus driver may deliberately slow down in smooth traffic so as to avoid early arrival at the downstream stop. Consequently, the lowered travel speed may leave the passengers with the impression that the bus service is inefficient. These two kinds of drivers' behavior are common in China [6, 7]. The root cause is that the initial timetables are usually nonoptimal considering the real-time traffic conditions. Therefore, the retrieved GPS data cannot be used directly. To obtain the actual travel speed and to further divide the operating time, the effect of drivers' behavior should be considered.

Scholars have conducted much research on the optimization of bus schedule schemes but have rarely investigated the division of the operating time [8–11]. To evaluate the effectiveness of a bus schedule scheme, Patnaik et al. [12] selected as indexes the numbers of passengers boarding and alighting the bus and the number of midway stops. The buses from the starting stop were then divided into several classes. The data used to develop the models were collected by the Automatic Passenger Counters (APC) on buses operated

by a transit agency in the northeast region of the United States. Guihaire and Hao [13] presented a global review of the crucial strategic and tactical steps of transit planning: the design and schedule of the network. They pointed out that the bus operating period mainly depended on the passengers' requirements which were different at different times. However, no analytic method has been developed for time of day partition. Using ridership data from a bus smart card system, Yue [14] obtained an ordered sampling of the passengers' arrival ratio curve and divided the operating time into multiple intervals using the Fisher optimal segmentation method. In his model, only the passenger volume was considered and the bus travel speed was neglected. As a result, the bus operating conditions were not fully considered during the partition. Shen et al. [15] proposed an improved K-means clustering algorithm for the division of the bus operating period based on GPS data. However, only the bus travel speed was used and the passenger demand was not considered. Given that, in different time intervals, a transit agency tends to arrange different bus dispatching frequencies because of the different passenger demand, this study becomes less practically promising. Bie et al. [16] selected the dwell time at each stop and the travel time between each pair of them as indexes and developed a rapid division algorithm. This is the first study that considers both the bus travel speed and the passenger demand in time of day partition. However, the GPS data were used directly without considering the deliberate speed-up or slow-down movement.

The existing methods for operating time division exhibit two shortcomings: (i) only the passenger flow volume is taken into account and (ii) data are obtained typically through manual work which consumes much manpower and many other resources. The method proposed in this study builds the relationship between time division and bus schedule scheme and successfully addresses these shortcomings.

The contributions of this study are twofold. Firstly, we develop a method to identify whether there is deliberate speed-up or slow-down movement of a bus. A recovery method is then established for calculating the real bus travel time based on raw GPS data. To the best of our knowledge, no research so far has investigated this kind of problem. Secondly, a sequential clustering algorithm is developed to partition the operating period into multiple intervals based on the recovered bus travel time and dwell time at stops.

The structure of this paper is organized as follows. In Section 2, a recognition method for bus operating state is first developed followed by a recovery method for the bus travel time. A discussion is provided as to why the recovery travel time and dwell time are selected as division indexes. In Section 3, a sequential sample clustering algorithm is proposed to divide the operating time into multiple time intervals using the recovered travel time and dwell time. Section 4 presents a real case study and Section 5 concludes the paper.

2. Development of the Operating Time Division Method

2.1. Recognition of the Bus Operating State. In this paper, unless stated otherwise, all time is measured in units of seconds. Let us assume that a bus i passes m stops in total during an operating period n. According to its timetable, the planned travel time of bus i from the mth stop to the $(m+1)$th stop is denoted as $T_i^n(m, m+1)$. The planned operating time T_i^n can be written as follows:

$$T_i^n = \sum_{m=1}^{M-1} T_i^n (m, m+1). \tag{1}$$

When traveling along a route, a bus usually passes through three different kinds of regions, namely, stops, road sections, and intersections. Therefore, the planned travel time of bus i from the mth stop to the $(m+1)$th stop can be further divided as follows:

$$\begin{aligned} T_i^n (m, m+1) = a_i^n (m, m+1) + b_i^n (m, m+1) \\ + c_i^n (m+1), \end{aligned} \tag{2}$$

where $a_i^n(m, m+1)$ denotes the travel time spent at road sections, $b_i^n(m, m+1)$ denotes the travel time spent at intersections, and $c_i^n(m+1)$ denotes the travel time spent at bus stops.

$a_i^n(m, m+1)$, $b_i^n(m, m+1)$, and $c_i^n(m+1)$ can be extracted from GPS data in combination with a geographic information system (GIS) map. The actual travel time of the bus i from the mth stop to the $(m+1)$th stop, denoted as $\widehat{T}_i^n(m, m+1)$, can be rewritten as follows:

$$\begin{aligned} \widehat{T}_i^n (m, m+1) = \widehat{a}_i^n (m, m+1) + \widehat{b}_i^n (m, m+1) \\ + \widehat{c}_i^n (m+1). \end{aligned} \tag{3}$$

(1) Recognition of a Driver's Deliberate Acceleration. Theoretically, the bus travel time at road sections and intersections increases under traffic jams.

$$\begin{aligned} \widehat{a}_i^n (m, m+1) &> a_i^n (m, m+1), \\ \widehat{b}_i^n (m, m+1) &> b_i^n (m, m+1), \\ \widehat{T}_i^n (m, m+1) &> T_i^n (m, m+1). \end{aligned} \tag{4}$$

At intersections, bus drivers tend to reduce speed because of the queuing vehicles and the restriction of changing lanes. However, a driver can frequently accelerate and decelerate at road sections to reduce the travel time and to ensure punctual arrivals at the downstream stops.

Case 1.

$$\widehat{T}_i^n (m, m+1) = T_i^n (m, m+1). \tag{5}$$

In Case 1, although a bus may be delayed at intersections, it still arrives at the downstream stops on time due to deliberate acceleration at road sections.

Case 2.

$$\begin{aligned}&\widehat{T}_i^n(m,m+1) > T_i^n(m,m+1),\\&\left[\widehat{T}_i^n(m,m+1) - T_i^n(m,m+1)\right]\\&\quad < \left[\widehat{b}_i^n(m,m+1) - b_i^n(m,m+1)\right]\\&\qquad + \left[\widehat{c}_i^n(m,m+1) - c_i^n(m,m+1)\right].\end{aligned} \tag{6}$$

In Case 2, although the driver may deliberately speed up the bus, it does not arrive on time at the downstream stops.

Case 3.

$$\begin{aligned}&\widehat{T}_i^n(m,m+1) > T_i^n(m,m+1),\\&\left[\widehat{T}_i^n(m,m+1) - T_i^n(m,m+1)\right]\\&\quad \geq \left[\widehat{b}_i^n(m,m+1) - b_i^n(m,m+1)\right]\\&\qquad + \left[\widehat{c}_i^n(m,m+1) - c_i^n(m,m+1)\right].\end{aligned} \tag{7}$$

In Case 3, the increase in the bus travel time at road sections exceeds or equals the total increase in the travel time spent at intersections and in the dwell time at stops. The bus may run normally or undergo deliberate acceleration.

(2) Recognition of a Driver's Deliberate Deceleration. When the traffic volume is low, the bus travel times at road sections and intersections may decline.

$$\begin{aligned}&\widehat{a}_i^n(m,m+1) < a_i^n(m,m+1),\\&\widehat{b}_i^n(m,m+1) < b_i^n(m,m+1),\\&\widehat{T}_i^n(m,m+1) < T_i^n(m,m+1).\end{aligned} \tag{8}$$

Theoretically, if the timetable is not optimized in real time, the driver may deliberately slow down the bus to enable punctual arrivals according to the schedule.

Case 1.

$$\widehat{T}_i^n(m,m+1) = T_i^n(m,m+1). \tag{9}$$

In Case 1, although the bus is slightly delayed at intersections, it still arrives at the downstream stops on time, since the driver deliberately slows down the bus.

Case 2.

$$\begin{aligned}&\widehat{T}_i^n(m,m+1) < T_i^n(m,m+1),\\&\left[T_i^n(m,m+1) - \widehat{T}_i^n(m,m+1)\right]\\&\quad < \left[b_i^n(m,m+1) - \widehat{b}_i^n(m,m+1)\right]\\&\qquad + \left[c_i^n(m,m+1) - \widehat{c}_i^n(m,m+1)\right].\end{aligned} \tag{10}$$

In Case 2, although the drive deliberately slows down the bus at road sections, the bus still arrives at the downstream stops ahead of the scheduled time.

Case 3.

$$\begin{aligned}&\widehat{T}_i^n(m,m+1) < T_i^n(m,m+1),\\&\left[T_i^n(m,m+1) - \widehat{T}_i^n(m,m+1)\right]\\&\quad \geq \left[b_i^n(m,m+1) - \widehat{b}_i^n(m,m+1)\right]\\&\qquad + \left[c_i^n(m,m+1) - \widehat{c}_i^n(m,m+1)\right].\end{aligned} \tag{11}$$

In Case 3, the decrease in the bus travel time at road sections exceeds or equals the total decrease in the travel time spent at intersections and in the dwell time at stops. The bus may run normally or undergo deliberate deceleration.

2.2. Recovery of the Optimal Travel Time on the Road. When a driver's deliberate acceleration or deceleration is recognized as discussed in Section 2.1, the retrieved GPS data cannot be directly used for the optimization of the schedule scheme. This effect should be considered for recovering the optimal bus travel time on the road.

The delay time of a bus at an intersection can be calculated by subtracting the travel time at a preset speed from the travel time spent at an intersection. During the operating period, n, a number of buses pass through the intersection and their average delay can be directly calculated. Assuming that $\overline{d}_i^1$ denotes the average delay at the timetable's initial operation stage, the traffic conditions will change after a certain period of time, and the average delay will become $\overline{d}_i^2$.

Generally speaking, the traffic flow on a road increases/decreases as a result of an increase/decrease in traffic flow at the adjacent intersection. According to the theory of traffic engineering, the travel time spent at a road section or at an intersection is directly proportional to the traffic flow. At a signalized intersection, the average delay $\overline{d}$ can be calculated by the following [17].

$$\begin{aligned}\overline{d} = {}& \frac{0.5C\left(1-\lambda_i\right)}{1-\left[\min\left(1,x_i\right)\cdot\lambda_i\right]}\\&+ 900T\left[\left(x_i-1\right)^2 + \sqrt{\left(x_i-1\right)^2 + \frac{4x_i}{\mathrm{Cap}_i\cdot T}}\right],\end{aligned} \tag{12}$$

where λ_i, x_i, and Cap_i denote the green ratio, degree of saturation, and traffic capacity, respectively, of the phase for bus i. T denotes the length of the analysis period and is generally set at 0.25 h.

$$x_i = \frac{q_i/S_i}{\lambda_i}, \tag{13}$$

where q_i and S_i denote the ratios of the arrival and saturation flows of the entrance lane for bus i, respectively.

For bus i, when the average delay changes from $\overline{d}_i^1$ to $\overline{d}_i^2$ while the other variables remain unchanged, the variation ratio of the flow at the entrance lane can be derived according to (12)-(13). Since r_i denotes the ratio of the flow after a certain period of time to the original one, r_i can also denote the

variation ratio of traffic flow which will be used for recovering the optimal travel time of the bus on the road.

Through field observations, a relationship is shown to exist between the average speed of traffic on urban roads and the flow. At low traffic flow, speed is insensitive to the increase in flow and only decreases slightly. When the flow increases and is close to the capacity of the road, the speed decreases significantly. When the flow is lower than the capacity of the road, the average speed varies with the flow in an approximate linear fashion:

$$q_i' = \alpha + \beta v_i, \tag{14}$$

where q_i' denotes the flow in pcu/h of the road section, v_i denotes the average speed of the traffic flow in km/h, and α and β are constants to be determined.

According to the characteristics of traffic flow, when free-flow speed u_f occurs, the traffic flow equals 0 ($q_i' = 0$). When the speed equals the optimal value u_m, the traffic flow q_i' reaches the maximum and the saturation flow ratio S_i is achieved. Therefore, the following equations hold:

$$\begin{aligned} \alpha + \beta \times u_f &= 0, \\ \alpha + \beta \times u_m &= S_i. \end{aligned} \tag{15}$$

By calculation, we can get $a = Su_f/(u_f - u_m)$, $b = S/(u_m - u_f)$.

$$v_i = \left(q_i' - \frac{S_i u_f}{u_f - u_m} \right) \frac{u_m - u_f}{S_i}. \tag{16}$$

Assuming that the flow changes to $r_i q_i'$ after the bus dispatching scheme is executed for a certain period of time, the average travel speed v_i' of the bus can be calculated by

$$v_i' = \left(r_i q_i' - \frac{S_i u_f}{u_f - u_m} \right) \frac{u_m - u_f}{S_i}. \tag{17}$$

Defining $r_i' = v_i'/v_i$, the following expressions can be obtained:

$$\begin{aligned} r_i' &= \frac{\left(u_f - u_m\right) r_i q_i' - S_i u_f}{\left(u_f - u_m\right) q_i' - S_i u_f}, \\ \widehat{a}^n(m, m+1) &= \frac{1}{r_i'} a^n(m, m+1). \end{aligned} \tag{18}$$

Let $a^n(m, m+1)$ denote the average travel time of the bus from the mth to the $(m+1)$th stop within the operating time period n at the timetable's initial operation stage. The optimal travel speed after a certain period of time becomes $\widehat{a}^n(m, m+1)$ which denotes the recovered average speed from the mth to the $(m+1)$th stop.

r_i' is the most important parameter which plays a decisive role in the travel time recovery process. Figure 1 illustrates the overall process for calculating r_i'.

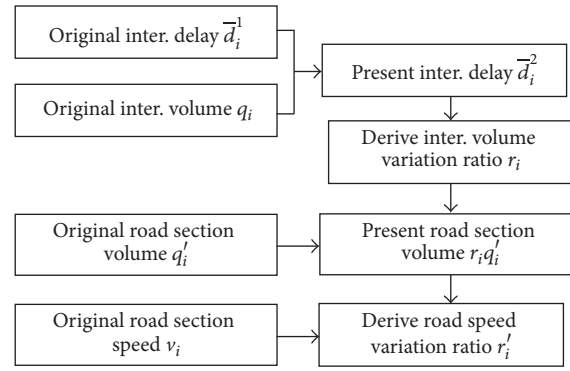

Figure 1: Flow chart of the bus travel time recovery.

2.3. Determination of the Operating Time Division Indexes. To divide the operating period, we first analyze all the historical data and then group into the same class buses with similar operating states and starting times into the same class. The corresponding operating time is referred to as a time interval. In this study, the dwell time at each stop and the interstop travel time (the recovered value as described in Section 2.2) are selected as the division indexes.

(1) Bus Dwell Time at Stops. The dispatching frequency affects the passenger volume of a bus route when the vehicle capacity of a bus is fixed. In each interval, the frequency is kept constant. Only when the passenger volume is also constant or slightly fluctuates will the passenger load factor of each dispatched bus be similar to one another. As a result, the uneven bus occupancy rate in an interval can be avoided.

With GPS data only, the number of alighting/boarding passengers at each stop is not available. However, empirical results show that the bus dwell time is positively proportional to the number of alighting/boarding passengers. Namely, a larger number of alighting/boarding passengers will result in longer dwell time. Though the bus dwell time is also affected by some other secondary factors such as the fare structure and the bus vehicle type (whether all doors can be used by the alighting passengers), they are all predetermined and remain unchanged for a given bus line. Hence the fluctuation of the bus dwell time at stops is mainly dependent on the number of alighting/boarding passengers. Therefore, the total dwell time at all stops is used to measure the passenger demand. Buses with similar total dwell time will be classified into the same time interval.

Let $\Delta D_{\max}$ denote the maximum permissible difference in the total dwell time at all stops for a bus in the same period. It can be calculated by $\Delta D_{\max} = \varepsilon \cdot \delta \cdot T_{ob}$ where δ denotes the maximum number of passengers, ε denotes the passenger carrying factor which is used mainly for adjusting the expected degree of crowdedness in the bus, and T_{ob} denotes the average boarding time of each passenger at each stop.

(2) Bus Interstop Travel Time. The bus operating status is affected not only by the arrival passenger volume at each stop, but also by the traffic conditions in real time. The

road traffic conditions influence the bus interstop travel time and hence the punctuality of the bus at each stop. For two buses dispatched consecutively from the same depot, if they have identical numbers of alighting/boarding passengers at each stop but different interstop travel time, they will experience different total travel time as well as different levels of punctuality at stops. Hence these two buses should not be classified into the same time of day interval.

Let $\Delta E_{\max}$ denote the maximum permissible difference in the total travel time among all stops for a bus in the same period. It can be calculated by $\Delta E_{\max} = \max\{0, H - \varepsilon \cdot \delta \cdot T_{ob}\}$, where H denotes the departure time interval for the buses as stated in the timetable.

Study [18] has analyzed the division indexes in different time of day for a bus route. However, the bus travel time obtained from GPS data was used directly. The deliberate acceleration or deceleration was not considered which renders the division results nonoptimal.

3. Operating Time Division Algorithm

Some classical clustering algorithms (such as K-means clustering) have achieved favorable results in index-based classification but are not suitable for this study. These classical algorithms do not take the order of data into account but quantify the correlation among data by using one of the distance metrics (such as the Euclidean and Mahalanobis distances). If the sequence of the buses is not taken into consideration, the buses with nonadjacent departure time intervals may be included in the same class. For example, when the first, second, third, fourth, tenth, and twentieth buses are included in the same operating period, this period can be divided into three subplots: subplot 1 includes the first, second, third, and fourth buses; subplot 2 includes the tenth bus; and subplot 3 includes the twentieth bus. Subplots 2 and 3 are quite short leading to frequent transitions between different bus dispatching schemes which reduces the management efficiency of the bus enterprise [19].

Given that the sequential sample clustering requires that the data sequence not be disturbed, a Fisher sequential sample clustering method (also referred to as optimal segmentation) is the most effective method [18]. There are 2^{n-1} division methods for n sequential samples. Each division method corresponds to segmentation. Among these segmentations, there exists an optimal segmentation that minimizes the difference within a segment and maximizes the difference among segments. To help achieve the optimal segmentation, the diameter of a class should be defined. After that the loss function is defined according to the constraint that the neighboring samples should be included in the same class. The optimal classification is found through a step-by-step recursive calculation with the objective of minimizing the loss function. The details of the procedure are described below.

(1) Calculation of the Diameter of a Class. In this study, the ordered variables are denoted as $x_1, x_2, \ldots, x_n$ (each variable x_i denotes an m-dimensional column vector, $i = 1, \ldots, n$). $m = 2$ given that the dwell time and the travel time are selected as two division indexes. Assuming that $\{x_i, x_{i+1}, \ldots, x_n\}$ denotes a segment ($1 \le i \le j \le n$), the diameter of a class (also referred to as the sum of the squares of deviation) $A(i, j)$ can be written as follows:

$$A(i, j) = \sum_{l=i}^{j} \left(x_l - \overline{x}_{i,j}\right)' \left(x_l - \overline{x}_{i,j}\right). \tag{19}$$

(2) Calculation of the Loss Function. For simplicity, the variable x_i ($i = 1, \ldots, n$) is denoted by its subscript i. Assuming that i_k denotes the first sample (vector) in the kth segment, the following method can be used for dividing the n ordered variables into K classes:

$$P(n, K) : \{i_1 = 1, i_1 + 1, \ldots, i_2 - 1\}, \\ \{i_2, i_2 + 1, \ldots, i_3 - 1\}, \ldots, \{i_K, i_K + 1, \ldots, n\}. \tag{20}$$

To use Fisher clustering, we need to define a loss function $e(P(n, K))$ to evaluate the quality of clustering. For a certain division method, the loss function $e(P(n, K))$ is defined as the sum of the squares of the deviations of all classes. Given n and K (the Fisher algorithm is applicable to cases with a known class number, K), the total sum of the squares of the deviations of all classes is fixed. Hence a smaller intraclass sum of squares and a larger interclass sum of squares give better classification results. In other words, clustering or segmentation aims to find a method which minimizes the loss function $e(P(n, K))$:

$$\text{Obj: } \min e(P(n, K)) = \min \sum_{k=1}^{K} A\left(i_k, i_{k+1} - 1\right). \tag{21}$$

To solve the above-described objective function, we use the following recursion:

$$\min e(P(n, K)) \\ = \min_{K \le i \le n} \{\min e(P(n - 1, K - 1)) + A(i, n)\}. \tag{22}$$

For example, when $K = 2$, $P^*(n, 2)$ is the optimal method among all possible division schemes that minimizes the loss function.

$$e\left(P^*(n, 2)\right) = \min e(P(n, 2)) \\ = \min_{2 \le i \le n} \{A(1, i - 1) + A(i, n)\}. \tag{23}$$

Using the method of induction, the recursion described in (23) can be derived which represents the optimal classification method of dividing n samples into K classes. It can be regarded as a combination of the optimal classification method of dividing $i - 1$ samples into $K - 1$ classes and the Kth segment which includes the remaining $n - i + 1$ samples.

There are two unique features of this algorithm. Firstly, it does not disturb the order of the dispatched buses. Hence the numbers of all buses that are classified into the same interval are adjacent. Secondly, the algorithm is not complex which takes less time to get the partition results and which can improve the computational efficiency.

(3) Final Division Based on Threshold Values of Two Indexes. By means of the above two steps, the dispatched buses are

Table 1: Time interval partition scheme and bus headway during the investigation period.

Interval number	Starting and ending times	Headway (min)	Bus quantity	Departure number
1	05:50–07:00	15	5	1–5
2	07:00–09:00	8	15	6–20
3	09:00–12:00	10	18	21–38
4	12:00–16:00	11	22	39–60
5	16:00–19:00	8	22	61–82
6	19:00–21:00	12	11	83–93

Table 2: Minimal lost function and starting codes of the last cluster in different partition methods.

	$K = 2$	$K = 3$	$K = 4$	$\cdots$	$K = 89$	$K = 90$	$K = 91$	$K = 92$
$n = 3$	0.0007 [2]			$\cdots$				
$n = 4$	0.05134 [3]	0.0437 [3]		$\cdots$				
$n = 5$	0.0094 [3]	0.0084 [3]	0.0008 [4]	$\cdots$				
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$n = 91$	9.121 [52]	8.068 [83]	7.482 [65]	$\cdots$	0.0026 [91]	0.0016 [91]		
$n = 92$	10.065 [52]	8.448 [83]	7.926 [65]	$\cdots$	0.0029 [92]	0.0026 [92]	0.0016 [92]	
$n = 93$	9.909 [48]	8.213 [82]	7.683 [72]	$\cdots$	0.0039 [93]	0.0029 [93]	0.0026 [93]	0.0016 [93]

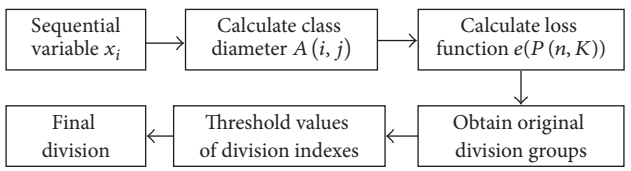

Figure 2: Flow chart of the division algorithm.

divided into K groups. However, it has not been determined whether the differences in the total dwell time and in the interstop travel time between two adjacent buses are smaller than the threshold values, which is thus evaluated in this step. In each group, if the differences in the two division indexes of two adjacent buses are larger than the threshold values, the two buses should be classified into different groups.

Figure 2 illustrates the overall process of the division algorithm.

4. Case Study

In this section, we apply the proposed time of day division method based on GPS data on the number 63 bus route in Harbin, China, as a case study.

4.1. Data Acquisition. Bus route 63 in Harbin has 21 stops in total. The line starts from Jiangong Community and goes all the way to Dajiang Community along the westbound direction. The operating distance of one direction is approximately 9.5 km. The bus enterprise has set the sampling interval of the GPS data at 30 seconds, which, however, cannot satisfy the requirement of this study. As a result, we carried out our own investigation of the bus line for two weeks (from Monday to Friday per week) during September 2013. In each bus, a GPS device was placed and connected to a laptop for real-time storage of the GPS data, which were later matched with a GIS map. Afterwards, the required travel time spent on road sections and intersections and the delay and dwell time at stops were extracted. During the investigation, the bus operating time was from 05:50 to 21:00. The operating period can be divided into 6 time intervals for each day. The specific starting and ending time points as well as the departure headways are listed in Table 1. In December 2014, we performed a second investigation for one week (Monday to Friday) and obtained the latest bus operating data. Compared with the first investigation, the total number of vehicles in Harbin had increased significantly. In addition, due to the winter snow on the road, vehicles moved more slowly and road congestion became even more serious. Acceleration and deceleration of the bus vehicles happened to be more frequent. As a result, the original schedule scheme was no longer suitable for the second investigation.

4.2. Division Results. Before the division of the operating time, the values of various parameters should be determined. δ and T_{ob} are constants which are set at 60 people per bus and 2.2 seconds per person, respectively. Given that $H = 8$ minutes, $H^a_{\max} = 2H$ and $H^a_{\min} = H$ and $\Delta D_{\max} = 112.2$ and $\Delta E_{\max} = 367.8$.

All the division indexes are normalized before used. Table 2 lists the minimal loss function based on the sequential clustering and the beginning label at the last time slot. The minimal loss function is calculated from the second column; that is, $K = 2$. The minimal loss functions of all the schemes of dividing the first i buses ($3 \le i \le 97$) into K classes are derived

to determine the optimal segmentation. Using $\min e(P(3,2))$ as an example, there are two division schemes which divide the first two buses into two classes, namely, ({1}, {2, 3}) and ({1, 2}, {3}).

$$\begin{aligned}\min e(P(3,2)) &= \min_{2\le j\le 3}\left(A_{1,j-1} + A_{j,3}\right)\\ &= \min\left[\left(A_{1,1} + A_{2,3}\right), \left(A_{1,2} + A_{3,3}\right)\right]\\ &= \min(0.0007, 0.0076) = 0.0007.\end{aligned} \quad (24)$$

The optimal segmentation is ({1}, {2, 3}), and the beginning label of the last class (i.e., 2) is recorded. As shown in the second row and the second column in Table 2, [2] on the right of 0.0007 represents the division of the first three buses into 2 classes where the beginning label of the second class is 2 and the corresponding minimal loss function is 0.0007. Moreover, the division indexes (the average dwell time at stops and the average travel time among stops) are different for different buses in a class which should be taken into account in the classification. The buses whose division indexes are smaller than the thresholds are grouped into the same class. For example, there are 92 division schemes when dividing 93 samples into 2 classes. Before the calculation of the loss function, we should first evaluate whether the thresholds $\Delta D_{\max}$ and $\Delta E_{\max}$ are satisfied and delete those division schemes that do not satisfy the requirement. Only after that can the loss functions of the remaining division schemes be calculated so as to determine the optimal division.

As shown in Table 2, the sequential clustering algorithm cannot determine the class number K but can only determine the optimal class number according to the variation in the minimal error function. It can be observed that, in the last row of Table 2 ($n = 93$), the minimal error function of 93 sample data decreases gradually with an increasing K. A greater K suggests a finer division and, accordingly, fewer buses are included in a class in which the difference is smaller. However, a bus enterprise does not necessarily want to increase the number of the operating time slots, since doing so will not only increase the frequency to update the dispatching schemes but also require more transition schemes between different dispatching schemes. Frequent transitions may reduce the operating efficiency of bus transit [7]. In studies [7, 8], the value of K was determined by the manager. For this study, with reference to the previous research, we consulted the administration department of the bus enterprise and finally set the value of K at 8; that is, the operating time of the number 63 bus is divided into 8 time intervals as shown in Table 3.

Table 3: Final partition results of operation time intervals for bus route 63.

Interval number	Starting and ending times	Interval number	Starting and ending times
1	5:50–7:05	5	12:49–15:01
2	7:05–8:09	6	15:01–16:39
3	8:09–11:05	7	16:39–19:15
4	11:05–12:49	8	19:15–21:00

5. Conclusion

This study first recovers the bus travel time on the road based on the historical GPS data and then divides the bus operating time using a sequential clustering algorithm. The main conclusions are as follows:

(1) The bus travel time data collected from the bus-mounted GPS cannot truly reflect the real operating state of the bus vehicle. Drivers' behavior should be taken into account for data correction.

(2) For the division of the operating time, the division algorithm is more sensitive to the threshold value of the dwell time at stops. A smaller threshold value may easily make the division finer.

(3) A sequential clustering method can ensure that the order of the adjacent buses is not disrupted in order to achieve a favorable division of the bus operating time.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Project no. 51578199). This work was performed at the Key Laboratory of Advanced Materials & Intelligent. Control Technology on Transportation Safety, Ministry of Communications, China.

References

[1] Z. Liu, Y. Yan, X. Qu, and Y. Zhang, "Bus stop-skipping scheme with random travel time," *Transportation Research C: Emerging Technologies*, vol. 35, pp. 46–56, 2013.

[2] X. Ma and Y. Wang, "Development of a data-driven platform for transit performance measures using smart card and GPS data," *Journal of Transportation Engineering*, vol. 140, no. 12, Article ID 04014063, 2014.

[3] X. Ma, Y. J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 1–12, 2013.

[4] X. Ma, C. Liu, H. Wen, Y. Wang, and Y. Wu, "Understanding commuting patterns using transit smart card data," *Journal of Transport Geography*, vol. 58, pp. 135–145, 2017.

[5] Y. Li, X. Wang, S. Sun, X. Ma, and G. Lu, "Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 306–328, 2017.

[6] X. Gong, X. Guo, X. Dou, and L. Lu, "Bus travel time deviation analysis using automatic vehicle location data and structural equation modeling," *Mathematical Problems in Engineering*, vol. 2015, Article ID 410234, 9 pages, 2015.

[7] M. Chen, X. Liu, and J. Xia, "Dynamic prediction method with schedule recovery impact for bus arrival time," *Transportation Research Record*, no. 1923, pp. 208–217, 2005.

[8] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 187–197, 2015.

[9] C. Ding, D. Wang, C. Liu, Y. Zhang, and J. Yang, "Exploring the influence of built environment on travel mode choice considering the mediating effects of car ownership and travel distance," *Transportation Research Part A: Policy and Practice*, vol. 100, pp. 65–80, 2017.

[10] C. Ding, X. Wu, G. Yu, and Y. Wang, "A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data," *Transportation Research Part C: Emerging Technologies*, vol. 72, pp. 225–238, 2016.

[11] C. Ding, S. Mishra, G. Lu, J. Yang, and C. Liu, "Influences of built environment characteristics and individual factors on commuting distance: a multilevel mixture hazard modeling approach," *Transportation Research Part D: Transport and Environment*, vol. 51, pp. 314–325, 2017.

[12] J. Patnaik, S. Chien, and A. Bladikas, "Using data mining techniques on apc data to develop effective bus scheduling plans," *Journal of Systemics, Cybernetics and Informatics*, vol. 4, no. 1, pp. 86–90, 2006.

[13] V. Guihaire and J. K. Hao, "Transit network design and scheduling: a global review," *Transportation Research A: Policy and Practice*, vol. 42, no. 10, pp. 1251–1273, 2008.

[14] D. Z. Yue, *Optimal timetable research based on passenger arrival rate*, School of Control Science and Engineering, Shandong University, Jinan, China, 2014 (Chinese).

[15] Y. D Shen, T. H Zhang, and J. Xu, "Homogeneous bus running time bands analysis based on K-means algorithms," *Journal of Transportation Systems Engineering and Information Technology*, vol. 14, no. 2, pp. 87–93, 2014 (Chinese).

[16] Y. Bie, X. Gong, and Z. Liu, "Time of day intervals partition for bus schedule using GPS data," *Transportation Research Part C*, vol. 60, pp. 443–456, 2015.

[17] Transportation Research Board, "Highway Capacity Manual," 2000.

[18] R. Guo and Y. Zhang, "Identifying time-of-day breakpoints based on nonintrusive data collection platforms," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 18, no. 2, pp. 164–174, 2014.

[19] P. Yu, H. Chi, and X. C Tan, "A study on flight altitude discrepancy base on the fisher ordinal samples cluster method," *Chinese Journal of Management Science*, vol. 18, no. 5, pp. 130–136, 2010.