GITSS2015

Unifying Time Reference of Smart Card Data Using Dynamic Time Warping

Haibo Li\textsuperscript{a,b}, Xuewu Chen\textsuperscript{a,b,}\textsuperscript{*}

\textsuperscript{a}Jiangsu Key Laboratory of Urban ITS, Southeast University, 2 Sipailou, Nanjing 210096, China
\textsuperscript{b}Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, 2 Sipailou, Nanjing 210096, China

Abstract

The wide use of smart card automatic fare collection (AFC) systems in public transit makes it popular to analyse public transit user behaviour based on smart card data. Most smart-card-based researches are time-related, but smart card data time is not always reliable because smart card systems are usually off-line and lack of maintenance in China, and smart card data of different buses often shares different time references. This paper explores the application of dynamic time warping to unify time reference of smart card data. Based on the analysis of the relationship between boarding time in smart card data and arrival-departure time in automatic vehicle location (AVL) data, a smart card data time reference unification algorithm was proposed. The results indicate that the algorithm can work out reliable time offsets by numerical calculation and runs well with integral automatic vehicle location data. Due to the fact that the time reference of smart card data for each bus changes daily, the best way to put an end to smart card time reference problem is to make smart card systems online, such as combining smart card readers with automatic vehicle location systems.

1. Introduction

During the past few decades, smart card technology, which was first developed by two German inventors Jürgen Dethloff and Helmut Grötrupp in 1968 [1], has been widely used in public transit automatic fare collection (AFC) systems all over the world. Although smart card automatic fare collection systems are originally designed for fare...
collection, their popularity makes it possible to gather an enormous amount of daily travel data that, if captured, archived, and analyzed properly, holds promise for improving service planning, scheduling, and performance analysis practices.

In recent years, many researches on the use of smart card data in public transit have been conducted, most of which focus on public transit user behavior, such as origin and destination inference, transfer point estimation and travel pattern analysis [2, 3].

In most AFC systems, especially in China, public transit users only need to swipe their smart cards when getting on a bus. As a result, smart card data records no detailed boarding information except for boarding time. Thanks to automatic vehicle location (AVL) systems, which record bus locations every few seconds, it’s possible to infer boarding location (origin) by matching boarding time in smart card data with arrival-departure time in AVL data, see Cui [4], Wang and Attanucci et al. [5]. The alighting location (destination) inferences are based on the trip-chaining method with the assumption that a public transit user would like to alight at the nearest stop to the boarding stop of the next trip [4, 6, 7].

Nowadays public transit agencies are placing more importance on transfers. With only smart card data, a method based on elapsed time threshold is often used to identify transfers, see Hofmann and Wilson et al. [8], Seaborn and Attanucci et al. [9]. If the elapsed time between a passenger’s two consequent boardings on different routes is less than the threshold given, the boardings are regarded as a transfer journey. But this method ignores the fact that the elapsed time varies widely with different travelling distances. Therefore, Li and Chen [10] redefined the elapsed time as waiting time at transfer stations by inferring alighting time and walking time based on AVL data.

In addition, it’s another major objective of using smart card data to analyze travel pattern of public transit users. Different classification techniques, such as K-means algorithm which is one of the most popular clustering algorithm [11], have been used to identify different user groups with similar pattern profiles. Many researches explore the similarities in travel patterns using boarding time as the main travel characteristic [12-14].

As mentioned above, recent researches have examined the potential benefits of using smart card data in public transit. But if smart card data is not so reliable, analysis based on smart card data will be fallacious and useless. We know that boarding time in smart card data is recorded according to the time reference of each smart card reader, most of which are offline and not well maintained in China. It leads to the result that smart card data collected by different smart card readers in the same AFC system have different time references. Therefore, it’s not surprising to find that the boarding time of the next trip may be earlier than that of the previous trip. This paper aims at unifying time reference of smart card data by using dynamic time warping.

The rest of the paper is organized as follows. First, it’s a brief introduction to smart card and AVL data. Next, a smart card data time reference unification algorithm based on dynamic time warping (DTW) is proposed. Then, we present some interesting results with discussions. Finally, this paper ends with concluding remarks and proposes a few suggestions for the future work.

2. Data

2.1. Smart card data

Smart card data is originally collected for financial analysis and its data structure varies with different pricing schemes. This paper focus on the data structure of the pay-per-trip pricing scheme, which is the most popular public transit pricing scheme in China. Different form distance-based AFC systems, public transit users pay a fixed fare for each trip in pay-per-trip AFC systems, regardless how long they actually travel. A pay-per-trip AFC system is often an entry-only system, in which public transit users only need to swipe their smart cards when getting on a bus. As a result, no alighting information, such as alighting time or alighting location, would be recorded in this kind of smart card data. Table 1 shows the contents of the information recorded in the smart card data used in this study, which includes card ID, card type, route ID, bus ID and boarding time. The dataset was obtained from Suzhou Municipal Transportation Bureau, Jiangsu province, China, and covers a one-week period from May 13 (Monday) to May 19 (Sunday), 2013.
Table 1. Information recorded in smart card data for pay-per-trip AFC systems.

<table>
<thead>
<tr>
<th>Information</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Card ID</td>
<td>Unique number for each smart card</td>
</tr>
<tr>
<td>Card type</td>
<td>Passenger groups, such as adult, student, and elder</td>
</tr>
<tr>
<td>Route ID</td>
<td>Given number of each route</td>
</tr>
<tr>
<td>Bus ID</td>
<td>Given number of each bus</td>
</tr>
<tr>
<td>Boarding time</td>
<td>Boarding time (year/month/day/hour/minute/second)</td>
</tr>
</tbody>
</table>

2.2. AVL data

Now more and more buses are equipped with AVL systems based on satellite positioning systems, and it becomes easy to collect bus location data of various precisions by different sampling intervals. Generally, there’re two kinds of AVL data, one records absolute locations in terms of latitude and longitude coordinates and the other one records relative locations, such as bus station names. Absolute location data contains more detailed information on bus running conditions, while relative location data focuses on spatial relations between buses and stations. With latitude and longitude coordinates of each bus station, it’s not difficult to convert absolute locations to relative ones. Table 2 shows the data structure of the AVL data used in this paper. It’s worth mentioning that AVL data shares the same time reference, which is synchronized with satellite time. As a result, AVL data is chosen here for smart card time reference unification.

Table 2. Data structure of AVL data (relative location data).

<table>
<thead>
<tr>
<th>Information</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus ID</td>
<td>Given number of each bus</td>
</tr>
<tr>
<td>Route ID</td>
<td>Given number of each route</td>
</tr>
<tr>
<td>Route direction</td>
<td>Running direction of the bus</td>
</tr>
<tr>
<td>Station name</td>
<td>Name of the bus station</td>
</tr>
<tr>
<td>Arrival time</td>
<td>Arrival time (year/month/day/hour/minute/second) at the bus station</td>
</tr>
<tr>
<td>Departure time</td>
<td>Departure time (year/month/day/hour/minute/second) at the bus station</td>
</tr>
</tbody>
</table>

3. Methodology

3.1. Basic Assumptions

It’s common sense that public transit users could get on a bus only when it stops at a bus station and no public transit user would like to get on a bus when it’s still moving, except in an emergency. Field observations show that public transit users tend to pay for their trips as soon as getting on buses and only a few would pay after a bus leaves a station because of some special reasons, like being elderly or disabled, taking large luggage, temporary failure to find smart card, etc. It means that boarding time in smart card data usually is between corresponding arrival and departure time in AVL data. This relationship is often applied as a rule to determine boarding locations of public transit users.

It’s easy to infer boarding locations for all smart card data with complete AVL data. However, time reference error of smart card data could have a negative influence on boarding location inference. As shown in Fig. 1, some boarding time would be later than corresponding departure time when the time reference of smart card data goes fast, or earlier than corresponding arrival time when the time reference of smart card data goes slow. In other words, time reference error could lead to some boardings unrecognizable. As we have learned from experience, the greater the time difference, the more the unrecognizable boardings.
According to the above analysis, we put forward three basic assumptions:

- Boarding time is always later than corresponding arrival time and earlier than corresponding departure time.
- The number of recognizable boardings reaches its peak when there’s no time difference between time references of smart card and AVL data.
- The optimal time difference between smart card and AVL data is the time difference which maximizes the number of recognizable boardings.

Based on the assumptions, we can search for the optimal time reference of smart card data by numerical calculation. But it’s common to find blank values for arrival or departure time in AVL data. To avoid possible negative influence, only arrival time is used for smart card time reference unification in this paper.

3.2. Dynamic Time Warping

Dynamic time warping (DTW), which was first introduced by Bellman and Kalaba [15], is a well-known algorithm applied in many areas. The algorithm can be used to find the optimal alignment between two time series if one time series may be “warped” non-linearly by stretching or shrinking it along its time axis.

The methodology for dynamic time warping is as follows [16-19]. Assume two time series, a sequence \( C \) of length \( m \), \( C = (c_1, c_2, \ldots, c_m) \) and a sequence \( A \) of length \( n \), \( A = (a_1, a_2, \ldots, a_n) \). The algorithm starts by building an \( m \)-by-\( n \) distance matrix, \( D(C, A) \), where each element \((i, j)\) represents the distance between the points \( c_i \) and \( a_j \), denoted as \( d(c_i, a_j) \), which can be any suitable distance function, such as:

\[
d(c_i, a_j) = |c_i - a_j|^p
\]  

A warping path, which is a sequence of points corresponding to an alignment between \( C \) and \( A \), can be represented as:

\[
w = (w_1, w_2, \ldots, w_l), \max(m, n) \leq l \leq m + n - 1
\]  

where \( w_k = (i_k, j_k), i_k \in [1, m], j_k \in [1, n], \) and \( k = 1, 2, \ldots, l \).
Fig. 2. Warping Matrix and Optimal Warping Path by DTW.

The shortest warping path corresponds to the optimal alignment, as shown in Fig. 2. Usually, a warping path needs to be subject to the following constraints, see Senin [17]:

- **Boundary condition:** \( w_1=(1,1) \) and \( w_l=(m,n) \). It means that the warping path should start at the first and end at the last points of the aligned sequences.
- **Monotonicity condition:** \( i_{k-1} \leq i_k \) and \( j_{k-1} \leq j_k \). This condition ensures that the path will not turn back on itself.
- **Step size condition:** \( w_{k-1}w_{k,1} \in \{(1,1),(1,0),(0,1)\} \). This prevents the warping path from long jump while aligning sequences.

Rabiner and Juang [20] introduced a dynamic programming algorithm to find the shortest warping path efficiently. The approach is based on the cumulative distance matrix, \( \gamma \), which defined as follows:

\[
\begin{align*}
\gamma(1, j) &= \sum_{k=1}^{j} d(1, k) \\
\gamma(i, 1) &= \sum_{i=1}^{i} d(k, 1) \\
\gamma(i, j) &= d(c_i, a_j) + \min\{\gamma(i-1, j), \gamma(i, j-1), \gamma(i-1, j-1)\} \\
\end{align*}
\]

Each element \((i, j)\) is the minimum cumulative distance of all the warping paths starting at \((1,1)\) and ending at \((i, j)\). Thus the minimum cumulated distance among all possible warping path, \(d^*_{w}(A, C) = (m, n)\).

This paper attempts to apply dynamic time warping to find the optimal alignment between boarding time series and arrival time series. Suppose sequence \( C \) stands for boarding time series and sequence \( A \) stands for arrival time series. As we known, it’s a one-to-many relationship between arrival time and boarding time, because two or more passengers can get on the same bus during the dwell time while a passenger can’t get on different buses at the same time. Because of this, the monotonicity condition should be reduced to \( i_{k-1} < i_k \) and \( j_{k-1} \leq j_k \), and the step size condition should be reduced to \( w_{k-1}w_{k,1} \in \{(1,1),(1,0)\} \). Moreover, not all bus stations have passengers, then the boundary condition should be reset as \( w_1=(1,j_s) \) and \( w_l=(m,j_e) \), where \( j_s, j_e \in [1,n] \).

### 3.3. Time reference unification algorithm

Based on the modified dynamic time warping, we put forward the following time reference unification algorithm for smart card data:
**Step 1** Generate the bus list according to the smart card data which needs time reference unification. Set the default range of the time difference, $\Delta t$, based on the maximum time offset ever observed. However, larger range doesn’t always mean better results. The maximum time difference in this paper is set to 30 minutes, and the initial value of the time difference, $\Delta t$, is -1800 s.

**Step 2** Read smart card data and AVL data of the $k$-th bus in the bus list from the database.

**Step 3** Correct the boarding time in the smart card data using the given time difference, $\Delta t$.

**Step 4** Apply the modified dynamic time warping to find the optimal alignment between the corrected boarding time series and arrival time series. To determine if the optimal alignment are reasonable, we define a statistical indicator called recognition rate, $R$, which can be calculated by:

$$R = \frac{n_{d(|c_i-a_j|)\leq \Delta}}{N_c}$$  \hspace{1cm} (4)

where $N_c$ is the total number of boardings in the smart card data, $n_{d(|c_i-a_j|)\leq \Delta}$ is the number of boardings meeting the condition: $d(c_i, a_j) = |c_i - a_j| \leq \Delta$, where $\Delta$ is the average dwell time and is set to $\Delta = 30$ s.

If $\Delta < 1800$, set $\Delta_t = \Delta_t + 1$ and go to step 3.

**Step 5** Find the maximum recognition rate, $R_{max}$, which is the optimal recognition rate. The time difference corresponding to the maximum recognition rate is the optimal time difference between the smart card and AVL data. To describe the difference between the maximum recognition rate and other recognition rates, we put forward an indicator called significance index, $SI$, which is defined as:

$$SI = \frac{R_{max} - \overline{R}}{S_R}$$  \hspace{1cm} (5)

where $\overline{R}$ is the average recognition rate of all assumed time differences, $S_R$ is the standard deviation of all recognition rates.

If the $k$-th bus is not the last one in the bus list, set $k = k + 1$ and go to step 2.

**Step 6** Return all optimal time differences and significance indexes.

### 4. Results and discussion

The results indicate that the assumptions made above are reasonable and the time reference unification algorithm is feasible and accurate, especially when both of smart card and AVL data is complete.

Taking the bus numbered 558307 for example, when the assumed time difference gets close to the optimal one, the recognition rate increases and the cumulated distance decreases on the whole, as shown in Fig. 3. The optimal time difference is supposed to appear at the highest peak of the TD-RR curve or the lowest valley of the TD-Distance curve or, where the recognition rate reaches its maximum value and the cumulated distance reaches its minimum value. According to the TD-RR curve, the optimal time difference of the bus numbered 558307 in May 25, 2013 is -102 s, which means that boarding time in smart card data is 102 seconds faster than the actual values.

According to the definition of recognition rate, the optimal recognition rate is sensitive to the integrity of AVL data. Due to system failures, the arrival time series of some buses are incomplete, which leads to low recognition rates. The optimal recognition rates of an average of 26.3 percent buses are less than 50 percent, while the optimal recognition rates of only an average of 22.9 percent buses are more than 80 percent. Low recognition rates usually imply AVL data loss, but don’t always mean unreliable optimal time differences. Through massive case studies, we find that different TD-RR curves of different recognition rates could have similar characteristics when they share similar significance indexes, no matter how many the optimal recognition rates are, even if the optimal recognition rates are less than 50 percent, as shown in Fig. 4. The optimal time differences corresponding to low recognition rates with high
significance indexes is often caused by discrete data loss, which has little influence on the optimal alignment between boarding time series and arrival time series.

Fig. 3. a) typical TD-RR curve; b) typical TD-Distance curve.

Fig. 5 gives a comparison of two TD-RR curves of the same high optimal recognition rate, more than 90 percent, but of different significance indexes. It indicates that high recognition rates must not always mean reliable optimal time differences. The rationality of the optimal time difference is also affected by the number of boardings in smart card data. Generally, the fewer the boardings, the less reliable the optimal time difference.

Fig. 6 indicates that the time references of smart card data change daily, and so are the recognition rates. To get correct smart card data, we should conduct the time reference unification every day. The mutable time references would make smart-card-based data analysis very difficult, because it not only adds a large amount of additional computation, but also makes smart card data unreliable.
Fig. 4. TD-RR curves of high, medium and low recognition rates.

Fig. 5. TD-RR curves of different significance indexes: a) SI=5.2; b) SI=2.7.
5. Conclusions

In recent years, there have been a lot of researches on the use of smart card data in public transit, but few pays attention to the possible time reference problem of smart card data, while many researches are time-related and inaccurate smart card data time could lead to unreliable or fallacious analysis.

This paper puts forward a time reference unification algorithm based on dynamic time warping to search the optimal time difference between smart card and AVL data by means of numerical calculation. The results show that the algorithm works very well when both smart card and AVL data are complete. Although small sample size of smart card data may have a negative impact on algorithm precision, the key of the time reference unification algorithm is to ensure the integrity of AVL data.

We also find that the time references of smart card data change daily. It means that the time reference unification should be conducted every day. It’s not only too much work, but also full of uncertainty. We may get an extremely unreliable time difference after complex computation and have to give up all the smart card data because of the failure to find the optimal time difference.

The time reference unification is necessary to archived smart card data, because time reference error has been integrated with smart card data and there is no other better choice. But if smart card readers are maintained properly,
the time reference problem can be avoided. However, the best way to put an end to the smart card time reference problem is to make smart card readers work online. For example, we can combine smart card systems with AVL systems, and let them share the same time reference. The system modification can not only solve the smart card time reference problem, but also make it possible to record boarding locations directly into smart card data.

Acknowledgements

This study is supported by the National Basic Research Program of China (No. 2012CB725402), the National Natural Science Foundation of China (No. 51178109) and the Jiangsu Province Ordinary University Graduate Research and Innovation Program (No. KYLX_0179). The authors also would like to acknowledge Suzhou Municipal Transportation Bureau for all data used in this study.

References