Extracting Bus Transit Boarding and Alighting Information Using Smart Card Transaction Data

Zhen Chen and Wei (David) Fan

USDOT Center for Advanced Multimodal Mobility Solutions and Education, Department of Civil and Environmental Engineering, University of North Carolina at Charlotte

Abstract

The smart card–based automated fare collection (AFC) technologies provide new opportunities for transportation data collection since smart card data (SCD) contain a significant amount of archived information that can be gathered and leveraged to help estimate public transit origin-destination (OD) matrices. Both boarding and alighting location detection are important particularly when there is no automatic vehicle location (AVL) system or other information in the database. With the analysis of raw data using limited information in this paper, an algorithm for boarding stop identification is presented after the trip direction and boarding cluster identification. Alighting stops are then identified based on the relationship between records. Finally, the feasibility and practicality of the methodology are tested using the bus transit SCD collected in Guangzhou, China.

Keywords: Transit smart card, automated fare collection, boarding location inference, alighting location inference

Introduction

As automated fare collection (AFC) systems have been widely implemented in the world today, smart card data (SCD) have played an important role in regional transportation system management. Specifically, information such as the transit rider origin-destination (OD) matrix and travel patterns can potentially be extracted from SCD. Considering the inconvenience of traditional travel survey methods, SCD-based research has become increasingly popular.

The studies on smart card data are broad. Pelletier, Trépanier, and Morency (2011) grouped the smart card data studies into three categories: strategic-level studies, tactical-level studies, and operational-level studies. Long, Sun, and Tao (2015) divided SCD relevant research into four categories: data processing and origin-destination
inference, transit system operation and management, spatial structure of city analysis, and mobility behavior and social networks analysis.

In general, the first step of SCD-based origin-destination estimation is to gather the boarding and alighting information. It is also important to note that in most cases, the smart card system is entry-swipe only, which may not have the alighting information of transit users. Sometimes the system may not even have the information about boarding stops/locations. As such, different research methods must be developed and applied to identify boarding and alighting information and build up the OD matrix. In this regard, two important and reasonable assumptions were made for this paper, as pointed out by Barry et al. (2002): (1) most riders would like to return to the destination station of their last trip to begin their next trip, and (2) the destination of most riders at the end of day is the origin stop where they begin their first trip of the day. Many SCD studies have been conducted to infer the origin-destination matrix based on these assumptions.

Identification of the boarding stop is usually the first step to SCD-based origin-destination inference. However, different cases may need to be considered since the data structure and supported database may vary. For example, Wang, Attanucci, and Wilson (2011) inferred trip origins with the help of both SCD and automatic vehicle location (AVL) data and achieved some results with high accuracy. When solving the origin inference problem without the AVL system, Barry, Freimer, and Slavin (2009) developed an approach for identifying specific boarding stops by utilizing scheduled run times to estimate the location of a bus along its route at the time of the automated fare collection (AFC) transaction. To detect the approximate bus boarding locations, they used the scheduled run time between stops along the route and adjusted the results with the transfer information obtained from smart card data. Yu, Deng, and Xiao (2009) adopted a hierarchical clustering method to classify the raw data based on the time interval between two adjacent transaction records. The referenced bus operation time was based on the fixed schedule and the criterion used was based on the difference between the cluster timestamp and the arrival time of each stop on the schedule. Ma et al. (2012) developed and utilized a Markov chain-based Bayesian decision tree algorithm to extract passengers’ origin information from the Beijing flat-bus AFC database. The authors introduced how to cluster transaction data by taking the following two steps: (1) sort transaction timestamps in ascending order, and (2) include the record as another cluster if the time difference between itself and the previous record was larger than 60 seconds and 30 minutes was set as the time threshold for a different trip. However, since the irregularity of the sample database and the time threshold configuration was a big issue, the direction identification task was not described in detail in this research. In short, it is still challenging to identify boarding stop information without other supported data sources.

Estimating the passenger origin-destination matrix is a major part of transportation planning study. Due to the inconvenience involved in traditional travel surveys, SCD-based research has become more popular recently. Over the past several decades, many automatic systems have been introduced in the public transit area, including smart card related automated fare collection (AFC), automatic vehicle location (AVL), and automatic passenger counting (APC). Many studies on destination information inference were conducted with the help of these supported databases (e.g., Cui 2006; Zhao, Rahbee, and Wilson 2007; Gordon et al. 2013). Specifically, historical smart card transaction time data provide a major data source for transit user destination information inference and have been widely used by both researchers and practitioners. Trépanier, Tranchant, and Chapleau (2007) conducted SCD-based destination estimation research with the help of 2003 Gatineau, Quebec, Smart Card Automated Fare Collection (SCAFC) data. To resolve the issue of limited transaction information about a “single trip” in which only one transaction record of a day existed, the authors tried to find a similar boarding
that occurred in previous days. Munizaga and Palma (2012) employed trip chaining methodology to build up the OD matrix. Several assumptions were made in this research, including, “If two transactions in a row are made in the Metro or along the same bus route, a destination is assumed between the two, regardless of the time interval, because it is very unlikely that someone would go out of the Metro network during a trip, unless he/she has something to do at that location,” and “If the time elapsed to the next transaction is over two hours, then this next transaction is the first segment of a new trip.” The trip chaining method and typical assumptions mentioned above were also utilized by several studies to infer alighting information (Nassir et al. 2011; Alsger et al. 2015).

The purpose of the research presented in this paper was to develop a systematic approach to illustrating how passenger boarding and alighting information could be mined from only the raw data derived from a smart card based automated fare collection (AFC) system, without the reference of AVL data, APC data, or any other supported database. Advanced methods were developed to acquire the bus transit operation direction information, passenger boarding cluster information, the specific boarding stop information, and the alighting stop information for each record. The methodology was tested using Guangzhou bus transit smart card data. It is important to note that effectively extracting boarding and alighting information can greatly assist in building up the OD matrix, which can then help decision makers plan, design, operate, and manage a more efficient public transit system. The contributions of this study include developing a systematic methodology that can be applied to most SCD systems and generating OD estimation results that can reveal bus transit users’ travel patterns based on the collected SCD, both of which will help transit operators and planning organizations develop objective and effective data-supported policies.

The remainder of this paper first discusses the structure and potential problems of smart card data, and then describes the direction identification algorithm. This is followed by the clustering of boarding activities and identification of boarding stops, as well as the alighting stop identification. Finally, the study’s numerical results are discussed in detail and conclusions are drawn accordingly.

**Methodology**

This section describes how to identify boarding and alighting stop information in detail. It first explains how to supplement the data when adequate information is lacking. Second, the relationship between records in the sample database is discussed and the boarding stop identification methodology is presented. With the help of boarding stop information, different scenarios are given based on the relationships between each record. Finally, the alighting stop identification algorithm is discussed.

**Data Description**

The study utilized SCD from the Guangzhou, China, transit system database as the raw data. A transaction record is generated each time a passenger boards: 100,000 transactions on four bus routes in five days were used for conducting relevant analyses.

The data recorded did not include the stop IDs where the passengers paid their fares or the alighting information. Table 1 shows a sample transaction record in the smart card system with the following information:
- **Route** – This field indicates which route the transit record belongs to.
- **Card ID** – This is a 16-digit number that uniquely identifies a smart card.
- **Card Type** – The card types in this study include general, student, disable, and senior cards.
- **Bus ID** – This field indicates on which bus the transaction takes place.
- **Transaction Time** – Timestamp information that is accurate to the second.

<table>
<thead>
<tr>
<th>Route</th>
<th>Card ID</th>
<th>Card Type</th>
<th>Bus ID</th>
<th>Transaction Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route A</td>
<td>99852761030</td>
<td>General</td>
<td>75</td>
<td>‘20150101001444</td>
</tr>
<tr>
<td>Route A</td>
<td>90009698490</td>
<td>General</td>
<td>75</td>
<td>‘20150101001505</td>
</tr>
<tr>
<td>Route A</td>
<td>95208436630</td>
<td>General</td>
<td>75</td>
<td>‘20150101001654</td>
</tr>
</tbody>
</table>

To ensure data quality, the data were first cleaned. A major problem with the data is related to consecutive swiping activities. Some travelers prefer to swipe their cards continuously, but records of the same card ID with too many continuous swipes include errors. Therefore, records with more than three continuous swipes were removed from the raw data (Chen and Fan 2018).

**Direction Information Identification**

Since the transaction database only recorded the transaction time and contained no direction information, the boarding stops could not be determined by directly matching transaction times with the transit operation schedule. To infer bus direction information, a direction labeling process was developed that used the following criteria as mentioned in a previous study (Chen and Fan 2018):

- "The direction will be changed when the time gap between the current transaction and last transaction is more than 30 minutes."
- "The time difference between current record and first record of last transaction sequence is close to the operation time on schedule."
- "The records with the transaction time before 6 a.m. on any day will be classified as belonging to the trip series of the previous day."

Figure 1 presents a flow chart that illustrates the details of the trip direction detection algorithm developed by Chen and Fan (2018).
FIGURE 1.
Trip direction identification algorithm
Transaction Data Clustering

Because several passengers usually board in an intensive period of time and multiple smart card swiping therefore occurs at one specific bus stop, it is convenient and more accurate to categorize several transaction records into specific boarding clusters based on the time interval between the transactions. Figure 2 presents a flow chart detailing information about the boarding cluster identification algorithm. The algorithm follows the criteria that the time interval threshold for two consecutive records is 60 seconds (Ma et al. 2012; Chen and Fan 2018). The records were categorized as the same boarding cluster if the interval was within 60 seconds; otherwise the boarding cluster was changed.

![Flow Chart of Boarding Cluster Identification Algorithm]

**FIGURE 2.**
Boarding cluster identification algorithm
Boarding Stop Information Extraction

With the help of boarding cluster information, the specific boarding stops can be inferred based on the difference in timestamps between adjacent boarding clusters. The boarding location of the first boarding cluster is assumed to be the first stop (i.e., origin) of the route. It should be noted that the bus operation time of most trips is shorter than the scheduled operation time of an entire bus trip. Hence, the boarding stop information extraction used the following rules (Chen and Fan 2018) to ensure that different clusters were assigned to different bus stops:

1. “The boarding stop of 1st boarding cluster is labeled as stop 1.”
2. “The difference between the timestamps of the 1st record of boarding cluster \(n+1\) and last record of boarding cluster \(n\) is \(\Delta t_{n+1}\), the boarding stop ID of boarding cluster \(n\) is \(i\); the operation time between station \(i\) and \(i+1\) is \(a_{i(i+1)}\) on the schedule.”
3. “If \(\Delta t_{n+1} < a_{i(i+1)}\), label the boarding stop as \(i+1\) to the records with boarding cluster \(n+1\).”
4. “If \(\Delta t_{n+1} > a_{i(i+1)}\), compare \(\Delta t_{n+1}\) and the operation time \(a_{i(i+2)}\) between station \(i\) and \(i+2\) on the schedule... until \(a_{i(i+k)} > \Delta t_{n+1}\), then label the boarding stop as \(i+k-1\) to the records with boarding cluster \(n+1\).”

Figure 3 presents the rules applied in the boarding stop information extraction process.

![Figure 3](image)

**FIGURE 3.**
Boarding information extraction

By applying the methodology as developed above to process the SCD collected from Guangzhou, China, the boarding information was identified successfully. Table 2 shows an example of the boarding stop identification results.
Extracting Bus Transit Boarding and Alighting Information Using Smart Card Transaction Data

TABLE 2.
An Example of the Boarding Stop Identification Results

<table>
<thead>
<tr>
<th>Route ID</th>
<th>Transaction ID</th>
<th>Card ID</th>
<th>Card Type</th>
<th>Bus ID</th>
<th>Transaction Date</th>
<th>Transaction Time</th>
<th>Direction</th>
<th>Boarding Cluster</th>
<th>Boarding Stop ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route A</td>
<td>10044</td>
<td>99852761030</td>
<td>Regular card</td>
<td>75</td>
<td>2015/1/1</td>
<td>0:14:44</td>
<td>B</td>
<td>1</td>
<td>1B</td>
</tr>
<tr>
<td>Route A</td>
<td>10061</td>
<td>90009698490</td>
<td>Regular card</td>
<td>75</td>
<td>2015/1/1</td>
<td>0:15:05</td>
<td>B</td>
<td>1</td>
<td>1B</td>
</tr>
<tr>
<td>Route A</td>
<td>10062</td>
<td>95208436630</td>
<td>Regular card</td>
<td>75</td>
<td>2015/1/1</td>
<td>0:16:54</td>
<td>B</td>
<td>2</td>
<td>2B</td>
</tr>
<tr>
<td>Route A</td>
<td>10063</td>
<td>99857026060</td>
<td>Regular card</td>
<td>75</td>
<td>2015/1/1</td>
<td>0:16:55</td>
<td>B</td>
<td>2</td>
<td>2B</td>
</tr>
<tr>
<td>Route A</td>
<td>2411</td>
<td>91005456310</td>
<td>Regular card</td>
<td>75</td>
<td>2015/1/1</td>
<td>0:17:31</td>
<td>B</td>
<td>2</td>
<td>2B</td>
</tr>
<tr>
<td>Route A</td>
<td>2412</td>
<td>99455451040</td>
<td>Regular card</td>
<td>75</td>
<td>2015/1/1</td>
<td>0:17:33</td>
<td>B</td>
<td>2</td>
<td>2B</td>
</tr>
<tr>
<td>Route A</td>
<td>2451</td>
<td>99460138890</td>
<td>Senior card</td>
<td>75</td>
<td>2015/1/1</td>
<td>0:27:04</td>
<td>B</td>
<td>3</td>
<td>5B</td>
</tr>
</tbody>
</table>

Alighting Stop Information Extraction

Based on the literature review and the raw data analysis, the alighting information for only parts of the records could be detected. Therefore, the transaction records for each card ID were categorized into different scenarios (Figure 4) based on the number of the records, timestamp of the records, and so on.
FIGURE 4.
Different scenarios of transaction records
An example of the records within each scenario is shown in Table 3. This step aims to further illustrate the identification methodology used for the alighting information.

**TABLE 3.**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Route ID</th>
<th>Transaction ID</th>
<th>Card ID</th>
<th>Card Type</th>
<th>Bus ID</th>
<th>Transaction Date</th>
<th>Transaction Time</th>
<th>Direction</th>
<th>Boarding Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>Route A</td>
<td>74</td>
<td>97305414700</td>
<td>Regular card</td>
<td>2400515</td>
<td>2015/1/1</td>
<td>19:49:50</td>
<td>A</td>
<td>2A</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>Route C</td>
<td>70666</td>
<td>95930288860</td>
<td>Senior card</td>
<td>90857004</td>
<td>2015/1/1</td>
<td>8:40:47</td>
<td>A</td>
<td>8A</td>
</tr>
<tr>
<td>Route C</td>
<td>55903</td>
<td>95930288860</td>
<td>Senior card</td>
<td>30580050</td>
<td>2015/1/1</td>
<td>9:14:01</td>
<td>A</td>
<td>11A</td>
<td></td>
</tr>
<tr>
<td>Scenario 3</td>
<td>Route A</td>
<td>23256</td>
<td>90000646530</td>
<td>Regular card</td>
<td>60582050</td>
<td>2015/1/2</td>
<td>18:29:31</td>
<td>A</td>
<td>7A</td>
</tr>
<tr>
<td>Route A</td>
<td>12146</td>
<td>90000646530</td>
<td>Regular card</td>
<td>5000005</td>
<td>2015/1/3</td>
<td>18:29:26</td>
<td>A</td>
<td>4A</td>
<td></td>
</tr>
<tr>
<td>Scenario 4</td>
<td>Route D</td>
<td>87812</td>
<td>98537901310</td>
<td>Regular card</td>
<td>60100304</td>
<td>2015/1/2</td>
<td>17:54:21</td>
<td>B</td>
<td>1B</td>
</tr>
<tr>
<td>Route D</td>
<td>91893</td>
<td>98537901310</td>
<td>Regular card</td>
<td>4000715</td>
<td>2015/1/2</td>
<td>19:43:01</td>
<td>B</td>
<td>11B</td>
<td></td>
</tr>
<tr>
<td>Route D</td>
<td>89104</td>
<td>98537901310</td>
<td>Regular card</td>
<td>7210197</td>
<td>2015/1/3</td>
<td>17:18:11</td>
<td>A</td>
<td>9A</td>
<td></td>
</tr>
<tr>
<td>Scenario 5</td>
<td>Route B</td>
<td>44512</td>
<td>92046091410</td>
<td>Regular card</td>
<td>40448004</td>
<td>2015/1/4</td>
<td>9:53:27</td>
<td>A</td>
<td>4A</td>
</tr>
<tr>
<td>Route B</td>
<td>49817</td>
<td>92046091410</td>
<td>Regular card</td>
<td>78004510</td>
<td>2015/1/4</td>
<td>18:39:43</td>
<td>B</td>
<td>12B</td>
<td></td>
</tr>
<tr>
<td>Scenario 6</td>
<td>Route B</td>
<td>36559</td>
<td>93737492690</td>
<td>Regular card</td>
<td>30547050</td>
<td>2015/1/2</td>
<td>22:16:59</td>
<td>A</td>
<td>7A</td>
</tr>
<tr>
<td>Route B</td>
<td>34173</td>
<td>93737492690</td>
<td>Regular card</td>
<td>78004510</td>
<td>2015/1/3</td>
<td>10:39:26</td>
<td>B</td>
<td>4B</td>
<td></td>
</tr>
<tr>
<td>Scenario 7</td>
<td>Route A</td>
<td>422</td>
<td>95092298010</td>
<td>Regular card</td>
<td>2008035</td>
<td>2015/1/1</td>
<td>11:57:43</td>
<td>A</td>
<td>1A</td>
</tr>
<tr>
<td>Route A</td>
<td>532</td>
<td>95092298010</td>
<td>Regular card</td>
<td>5000005</td>
<td>2015/1/1</td>
<td>12:17:19</td>
<td>A</td>
<td>8A</td>
<td></td>
</tr>
<tr>
<td>Route A</td>
<td>6277</td>
<td>95092298010</td>
<td>Regular card</td>
<td>5500290</td>
<td>2015/1/1</td>
<td>14:55:24</td>
<td>B</td>
<td>3B</td>
<td></td>
</tr>
</tbody>
</table>

**Scenario 1 – Single Trip:** Under this scenario, there is only one transaction record of the specific card ID. Based on the boarding information identification methodology, the trip origin (boarding stop) of this record can be inferred. However, the destination information is not recorded, and detection of the alighting information for such records does not seem possible. Therefore, the records are categorized as a single trip scenario and no alighting information is identified.

**Scenario 2 – Single Trip Chain:** Under this scenario, there are multiple records of the specific card ID and the transaction date and period of all the records are the same. The direction of these records is also the same. Based on the boarding information identification methodology as described, the trip origin (boarding stop) can be inferred. Based on the trip chaining theory, the alighting information for the first few records can also be detected, but not for the last record. Since the alighting information for the last record cannot be detected, the records are categorized as a single trip chain scenario.
Scenario 3 – Multiple Single Trips in Same Period: Under this scenario, there are multiple records of the specific card ID. Based on the boarding information identification methodology, the trip origin (boarding stop) can be inferred. However, all the records occurred in a same time period (AM or PM period) on different days and the alighting information still cannot be detected. The records under this scenario are categorized as multiple single trips in different days with no alighting information identified.

Scenario 4 – Single Trip and Trip Chain in Same Period: Under this scenario, there are multiple records of the specific card ID. All the records happened in the same period (AM or PM) and the directions of these records are the same. Based on the boarding information identification methodology, the trip origin (boarding stop) can be inferred. The alighting information for the first few records in the trip chain can be detected, but not for the other records. The records under this scenario are categorized as multiple single trip and trip chain in the same period.

Scenario 5 – Trips with Different Directions: Under this scenario, there are multiple records on the same day of the specific card ID. The directions of these records are different. Based on the boarding information identification methodology, the trip origin (boarding stop) can be inferred. The boarding stop is the potential alighting stop of another record.

Scenario 6 – Commute Trips on Different Days: Under this scenario, there are multiple records of the specific card ID that occurred on different days during both AM and PM time periods. The direction of the records is also different. Based on the boarding information identification methodology, the trip origin (boarding stop) can be inferred. The potential alighting information can also be mined for further analysis.

Scenario 7 – Commute Trips with Trip Chain: Under this scenario, there are multiple records of the specific card ID that happened on different days. The direction of the records in different periods is different. Based on the boarding information identification methodology, the trip origin (boarding stop) can be inferred. The potential alighting information under this scenario can be mined for further analysis.

As mentioned, the alighting information is necessary to detect transit riders’ origin-destination information. In order to identify alighting information, Barry’s assumptions (Barry et al. 2002) were used to help mine the potential information from the inadequate database. Based on the analysis of the seven scenarios, the records could be matched as Commute according to the following criteria:

- Same Card ID, same Transaction Date, and different Direction
- Same Card ID, different Transaction Date, different Time Period, and different Direction

The criteria for the records matched as Transfer include the following:

- Same Card ID, same Transaction Date, same Direction or different Route, and same Time Period

For all other cases, the records were described as No Matching.

Based on the transaction record matching results and the scenarios discussed, the alighting stop information on each record could be mined. For the Transfer matching records, the alighting stop identification follows the trip chaining theory mentioned in the literature review section. For the Commute matching records, if any record has only one match, it can be inferred that the boarding stop of one record is the alighting stop of the other. For a record with multiple matches, the selection criteria are listed below.
First level of selection – Select the matched records with the transaction that occurred on the same day. Based on the assumption mentioned before, such passengers would most likely go back to the origin of the first trip of the day.

Second level of selection – If there is no matched record occurring on the same day, then select the record with the transaction that happened during peak hour, since the activity during peak hour more likely exhibits the regular patterns.

Third level of selection – If there is no matched record on the same day and during peak hour, then select the record with the boarding stop that is nearest to the terminal station.

Based on the record matching results and the selection results discussed in this section, the alighting stop information identification method was applied to the study case. Table 4 presents a sample of the alighting stop identification results and includes a new column with the alighting stop identification results of each record.

### TABLE 4.
An Example of the Alighting Stop Identification Results

<table>
<thead>
<tr>
<th>Route ID</th>
<th>Transaction ID</th>
<th>Card ID</th>
<th>Bus ID</th>
<th>Transaction Date</th>
<th>Transaction Time</th>
<th>Direction</th>
<th>Boarding Stop</th>
<th>Alighting Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route A</td>
<td>15</td>
<td>90000434960</td>
<td>5500490</td>
<td>2015/1/2</td>
<td>9:54:18</td>
<td>A</td>
<td>4A</td>
<td>22B</td>
</tr>
<tr>
<td>Route A</td>
<td>16</td>
<td>90000434960</td>
<td>9500496</td>
<td>2015/1/2</td>
<td>15:39:25</td>
<td>B</td>
<td>22B</td>
<td>4A</td>
</tr>
<tr>
<td>Route B</td>
<td>17</td>
<td>98497328700</td>
<td>9800193</td>
<td>2015/1/1</td>
<td>17:48:29</td>
<td>B</td>
<td>4B</td>
<td>1A</td>
</tr>
<tr>
<td>Route B</td>
<td>18</td>
<td>98497328700</td>
<td>75063070</td>
<td>2015/1/2</td>
<td>6:53:30</td>
<td>A</td>
<td>1A</td>
<td>4B</td>
</tr>
<tr>
<td>Route B</td>
<td>19</td>
<td>98497328700</td>
<td>75063070</td>
<td>2015/1/3</td>
<td>6:49:58</td>
<td>A</td>
<td>1A</td>
<td>4B</td>
</tr>
<tr>
<td>Route B</td>
<td>20</td>
<td>98497328700</td>
<td>9800193</td>
<td>2015/1/4</td>
<td>6:50:57</td>
<td>A</td>
<td>1A</td>
<td>4B</td>
</tr>
<tr>
<td>Route D</td>
<td>21</td>
<td>98509428970</td>
<td>70847004</td>
<td>2015/1/1</td>
<td>7:43:21</td>
<td>A</td>
<td>2A</td>
<td>11B</td>
</tr>
<tr>
<td>Route D</td>
<td>22</td>
<td>98509428970</td>
<td>60100304</td>
<td>2015/1/2</td>
<td>9:18:52</td>
<td>A</td>
<td>3A</td>
<td>11B</td>
</tr>
<tr>
<td>Route D</td>
<td>23</td>
<td>98509428970</td>
<td>4210117</td>
<td>2015/1/2</td>
<td>21:56:45</td>
<td>A</td>
<td>16A</td>
<td>N/A</td>
</tr>
<tr>
<td>Route D</td>
<td>24</td>
<td>98509428970</td>
<td>77008740</td>
<td>2015/1/4</td>
<td>22:07:06</td>
<td>B</td>
<td>11B</td>
<td>2A</td>
</tr>
</tbody>
</table>

### Numerical Results

By applying the algorithms developed, the boarding and alighting locations were successfully identified based on the smart card data collected from Guangzhou, China, in this study. The key information gathered can be summarized as follows:

- **Boarding cluster information** – With the help of boarding cluster information, the transaction records for a whole trip along a route were divided into different cluster groups to identify the transactions that occurred at different stops. Boarding cluster information can greatly reduce potential errors and make it easier to identify boarding stop information.

- **Boarding stop information** – Boarding stop information was one of the key findings in this research. Additional information about passenger boarding counts can be mined and analyzed.

- **Alighting stop information** – Finally, the results show the estimated alighting stop of each transaction record, which will be helpful for estimating the origin-destination matrix.
Comparison of Average Passenger Counts during Each Time Period

The passenger boarding counts during each period are shown in Figure 5. These results indicate that the crest value of boarding activity occurred during different peak hours. For the AM period, the highest passenger counts occurred during 8–9 a.m. and the top three periods included 8–9 a.m., 10–11 a.m., and 9–10 a.m. For the PM period, the highest passenger counts occurred during 5–6 p.m. and the top three periods included 5–6 p.m., 6–7 p.m., and 4–5 p.m. This could be attributed to most government institutions and enterprises beginning their work around 9 a.m. and finishing around 6 p.m. Therefore, the travel patterns of the citizens in Guangzhou follow the daily work schedule, which is consistent with previous studies (Liu et al. 2009; Tao, Rohde, and Corcoran 2014).

FIGURE 5.
Passenger boarding counts during each time period

Frequency of Passengers’ Boarding Activities at Each Stop

Based on the boarding stop identification results, it is possible to estimate the frequency of boarding activities at each stop. The results shown in Figure 6 illustrate that most boarding activities occurred at the first several stops, and the passenger boarding counts decreased as the bus stops got closer to the terminal. Passenger boarding activities rarely occurred at the last several stops. The reason for the high passenger volume at the first stop could be the influence of the assumption of labeling the boarding stop of the first boarding cluster as stop 1. The decreasing trend at the last several stops is also consistent with previous studies (e.g., Liu et al. 2009). This clearly indicates both the travel habits of local passengers and the passenger boarding location characteristics from the traffic network viewpoint.
Analysis of Trip Characteristics of Different User Groups

Based on the boarding stop identification results, it is also possible to analyze the trip characteristics of each group of users. Figure 7 presents the passenger boarding counts of the different user groups on each route. The frequency distribution reveals that regular card users demonstrated notable peak-hour patterns, which is consistent with previous studies (e.g., Tao, Rohde, and Corcoran 2014). In comparison with regular card users, seniors displayed no pronounced peak-hour boarding patterns. However, the peak period of 9–10 a.m. in the figure is consistent with previous studies (e.g., Tao, Rohde, and Corcoran 2014). In contrast with the boarding pattern of regular card users, students had pronounced multi-peak period patterns throughout the day and the highest boarding frequency during the PM peak hour (18–20 p.m.). That could be the result of students having a tight daily schedule similar to regular card users. However, some students have more travel activities at noon and require more mobility than regular card users, which makes the proportion of their activity at noon higher than regular card users. In contrast with regular card users, the disable card users have no pronounced peak-hour boarding patterns throughout the day.
Conclusions

This research aimed to develop a methodology to extract boarding and alighting stop information by using the available transaction records with only basic route information, transaction time, and a few transfer activities, but without AVL, APC, and survey data. To reduce errors, the algorithm processed the data in the order of identification of direction information, boarding clusters, boarding stops, and alighting stops. Based on the acquired boarding information, the passenger count results were also derived, and they reveal that the crest values of boarding activity occurred during different peak hours. For the AM period, the highest volume occurred during 8–9 a.m. and the top three periods included 8–9 a.m., 10–11 a.m., and 9–10 a.m. For the PM period, the highest volume occurred during 5–6 p.m. and the top three periods included 5–6 p.m., 6–7 p.m., and 4–5 p.m. Most boarding activities occurred at the first several stops, and the passenger boarding counts decreased as the bus stops got closer to the terminal. The frequency distribution of each user group reveals that only regular card users demonstrated notable peak-hour patterns.

Several improvements could be made in the future, including consideration of the direction changing time threshold. Thirty minutes could be further investigated as the fixed time threshold utilized in this study. More scenarios will also need to be discovered and analyzed as more transit data become available.

The methodology and results of this study can be helpful for related work on the origin-destination estimation in the real world. In the future, the transfer activities of the whole transit network could be mined from the database if it contains information on more routes and for longer periods. The cluster analysis can also be
conducted to reveal passengers’ travel patterns if the survey data, card type, and land use data are available. Furthermore, the inferred boarding and alighting stop identification algorithm can be validated if the ground truth data are available.

Acknowledgments

The authors want to express their deepest gratitude to the financial support of the U.S. Department of Transportation, University Transportation Center, through the Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE) at the University of North Carolina at Charlotte (Grant Number: 69A3551747133). The authors also would like to thank the www.openits.cn website for providing the data for this research.

References


Extracting Bus Transit Boarding and Alighting Information Using Smart Card Transaction Data


About the Authors

Zhen Chen (zchen28@uncc.edu) is a third-year PhD student at the University of North Carolina at Charlotte.

Wei (David) Fan (wfan7@uncc.edu) currently serves as a full professor in the Department of Civil and Environmental Engineering at the University of North Carolina at Charlotte. He is the director of the USDOT University Transportation Center for Advanced Multimodal Mobility Solutions and Education. Dr. Fan holds a PhD in Civil Engineering – Transportation from the University of Texas at Austin (Hook ’em Horns!).