Transit Stop-Level Origin–Destination Estimation Through Use of Transit Schedule and Automated Data Collection System

Neema Nassir, Alireza Khani, Sang Gu Lee, Hyunsoo Noh, and Mark Hickman

As fare and data collection technology has developed, the resolution of collected data has reached the level of the individual traveler in investigations of transit passenger behavior. This paper investigates the use of these data to estimate passenger origins and destinations at the level of individual stops. Because of a lack of information from the fare collection system, researchers still need some estimate of passengers' alighting stops to complete each passenger trip chain on a specific day. Automated fare collection (AFC) and automated vehicle location (AVL) systems are the inputs to the estimation. Instead of typical AVL data, the paper proposes two models to estimate the alighting stop; both consider passenger trip chaining by using AFC data, transit schedule data (Google's General Transit Feed Specification), and automated passenger counter (APC) data. The paper validates the model by comparing the output to APC data with vehicle location data (APC-VL) and performs sensitivity analyses on several parameters in the models. To detect transfer trips, the new models propose a submodel that takes into account the effect of service headway in addition to some typical transfer time thresholds. Another contribution of this study is the relative relaxation of the search in finding the boarding stops, which enables the alternative algorithm to detect and fix possible errors in identification of the boarding stop for a transaction. As a result, the paper provides algorithms for the proposed models and sensitivity analysis for several predefined scenarios. The results are based on data and observed bus passenger behavior in the Minneapolis–Saint Paul, Minnesota, area.

Transit automated data collection (ADC) systems have allowed estimation of valuable behavioral patterns, especially for multimodal transit and with consideration of the sequence of passenger trips. Mainly, the interplay with data from the ADC systems—automatic fare collection (AFC), automated vehicle location (AVL), and other geographic information systems—provides more access to individual passenger's trip chain beyond that imagined at a more-aggregate level. Studies to identify passengers' trip sequences have been expanded to include multimodal transit networks as well as much larger networks (*1–6*). These studies have used this new technology with various methodologies to capture the individual trips.

In analysis of an individual's travel behavior by using ADC, generally, the main objective is to find the sequence of the passenger's trip from the origin stop to the destination stop (or origin to destination). But the information from AFC systems is still limited in its ability to infer the passenger's full sequence of trips. This limitation derives from the type of fare collection system, either closed or open (*2*). For example, a rail transit system may have a closed system that requires the use of a fare card at the origin and the destination but that may not require its use for internal transfers. In contrast, an open bus fare collection system may require the passenger to use a fare card only at boarding, not at alighting. Typically, open systems have been the main research interest in the development of a traveler's trip chain because the closed system provides both origin and destination (O-D) information of a trip.

To complete an individual's sequence of trips, the given ADC information requires the use of appropriate inferences. The main research using AVL and AFC data has been in estimating a passenger's reasonable alighting stop. In an open system, the principal inference comes in generating the connections for a sequence of trips by each card holder. If one assumes an inconvenience to walking, a frequent approach is to estimate the nearest alighting stop from the next AFC (boarding) transaction point on a trip (*1, 4, 6–10*). This estimate requires some inferences in the passengers' trip chaining, although ADC does not usually provide information on the cardholder's travel purposes, preferences, or attitudes (*11*). In addition, to generate the alighting stop alternatives, various thresholds for walking distances or travel times are used. In particular, Trépanier et al. introduced a methodology to approximate the nearest alighting point within the threshold of a 2-km (1.24-mi) Euclidean distance (*10*). Others (*7–9*) estimate the nearest alighting stop by considering the proximity of the arrival time of the transit vehicle (or run) to the next boarding time as well as the Euclidean distance to the nearest stop.

Another possible method is to conjecture whether an activity happens between two successive fare transactions. Several works (*7–9*) examine the relationship between the passenger's trip and activity occurrence by using AFC and AVL data. Because many fare collection systems allow different restrictions on the time available for a transfer, it is possible to have some simple, but relatively long, activity occur within the allowed transfer time. The easy way to

Department of Civil Engineering and Engineering Mechanics, University of Arizona, 1209 East 2nd Street, P.O. Box 210072, Tucson, AZ 85721-0072. Corresponding author: M. Hickman, mhickman@email.arizona.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2263, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 140–150.
DOI: 10.3141/2263-16

determine such an activity is to set up a transfer time boundary. Seaborn et al. examine the threshold of transfer time compared with an activity time for multimodal travel (rail and bus) in London (*3*). This decision of whether a time gap is a transfer or a simple activity is important because it directly affects the inferred O-D of the passenger trip.

To estimate a consistent alighting stop in a passenger's sequence of boarding transactions, two models are proposed. These models and previous ones are different in three major ways. First, for their inputs, the new models use Google's General Transit Feed Specification (GTFS), AFC, and automated passenger counter with vehicle location (APC-VL) data for the metro transit bus system in the Minneapolis–Saint Paul, Minnesota, area. This method uses GTFS and APC-VL data instead of AVL data. Today, it is relatively easy to obtain transit schedules because many transit agencies provide publicly available schedule information through GTFS. Second, this method enhances the ability of the decision process to understand AFC transactions, which consist of two types (called use types here): initial (beginning the first leg of the trip) and transfer. For understanding of the use type, a transfer time threshold has been applied in previous studies, such as a 40-min transfer time. However, the threshold could also be affected by the frequency of service (or headway) on the connecting route. By modeling the relationship between the transfer time threshold and headway, it is possible to provide a better inference for deciding whether an activity occurs. Third, another possible inference is applied as an alternative model in this study. In this model, a reliable alighting stop is estimated by relaxing the spatial search for boarding stops, to include the stops in the opposite direction, and then by trying to match the alighting stop from this alternative boarding stop.

The remaining sections of this paper are organized as follows. First comes a description of the data and preparation for analysis, by using AFC, APC-VL, and GTFS data. Then, the methodology by which the boarding and alighting stops are inferred is presented, and the use type for each transaction is estimated. In addition, an alternative algorithm is presented that considers the direction of service in matching boarding and alighting stops. Subsequently, detailed results of sensitivity analyses on model parameters are provided. Finally, concluding remarks and suggestions for future research are provided.

DATA: MINNEAPOLIS–SAINT PAUL METROPOLITAN TRANSIT

Data Description

The data in this research were obtained from Metro Transit operating in the Minneapolis–Saint Paul (Twin Cities) area and were excerpted from one month of data (November 2008). At the time, Metro Transit operated a fleet of 1,010 buses over 186 routes. The majority of bus headways ranged from 15 to 60 min. Less than 10% (18) of the routes had the minimum headways, ranging between 5 and 10 min, only during peak hours. The proportion of fare card users among all transit passengers is roughly 50%; this proportion was determine by comparing AFC records with boarding counts from the APC-VL data. Figure 1 shows the hourly distribution of total transactions on the basis of transaction date and time from AFC data. This graph manifests a conventional peaking pattern, with a huge percentage of the total transactions made during the morning (6:00 to 9:00 a.m.) and afternoon peak (4:00 to 7:00 p.m.) periods.

Table 1 presents the data recorded in the AFC, APC, and GTFS data sets. A more detailed description of the data follows.

AFC Data: Go-To Card

In the AFC system, a record is generated every time a user boards a bus. Each record has basic operational information, like the transaction date and time, route number, use type, fare type, bus identification (ID), run ID, and current location. In November 2008, the AFC transactions (2.17 million records) were made by 79,775 fare cards [identified by special serial number (SSN)]. Each SSN is considered an individual traveler because it is uniquely assigned to each Go-To card.

APC-VL Data

Stop-level boarding and alighting counts were collected from about 30% of operated buses, which were equipped with passenger counters. APC data (3.4 million records) also provided vehicle location information with the stop ID when boarding–alighting activities were

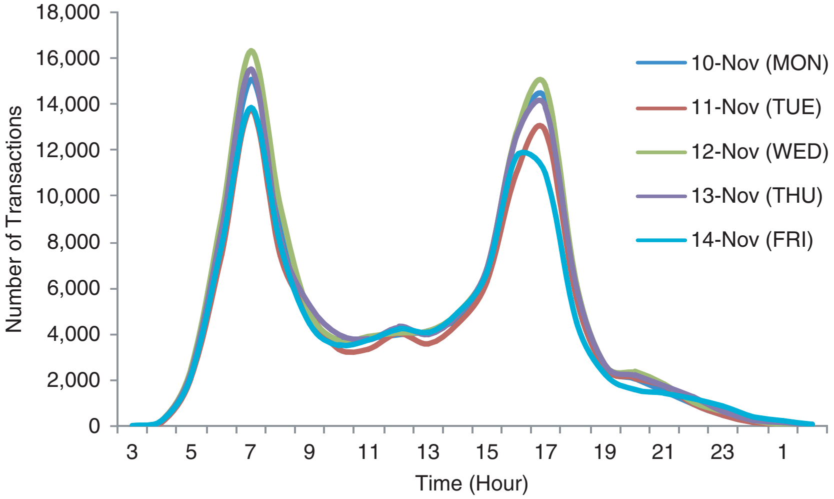

FIGURE 1 Hourly transaction distribution, week of November 10, 2008.

TABLE 1 Description of Each Data Set

Information	Description
AFC data	
Special serial number	Unique number of each go-to card
Transaction date	Boarding date
Transaction time	Boarding time
Route number	Given number of every bus route
Use type	Status of boarding (entry, refund, transfer)
Bus ID	Given number of every operated bus
LAT LONG data	Latitude–longitude of boarding location
APC data with vehicle location	
Vehicle ID	Given number of every operated bus (bus ID)
Time bracket start	Scheduled departure time at the first stop of trip
Time bracket end	Scheduled arrival time at the last stop of trip
Trip number	Given number of any given trip
Line ID	Given number of every bus route (route number)
Line direction	Directional information of any given trip
Stop sequence number	Given number of stop sequence for any given trip
Site ID	Given number of bus stop (stop ID)
Site latitude	Latitude of bus stop from vehicle location
Site longitude	Longitude of bus stop from vehicle location
Stops	
Stop ID	Given number of bus stop (site ID)
Stop name	Name of bus stop
Stop description	Direction and location of bus stop
Stop latitude	Latitude of bus stop
Stop longitude	Longitude of bus stop
Trips	
Route number	Given number of every bus route
Trip ID	Given number of every trip
Routes	
Route ID	Given number of every bus route
Calendar	
Service ID	Days of week when service is available
Stop times (schedule)	
Trip ID	Given number of every trip
Arrival time	Scheduled arrival time
Departure time	Scheduled departure time
Stop ID	Given number of bus stop (site ID)
Stop sequence number	Given stop sequence of bus trip

observed. In addition, the scheduled time and actual arrival–departure times at an individual stop were recorded in the APC data.

GTFS Data

GTFS is an open-source transit service package (*12, 13*) produced by hundreds of transit agencies in the United States. GTFS is typically presented as a series of text files (stops, stop times, routes, calendar, trips, etc.) with comma-separated values. The main advantage of using GTFS is to access the detailed schedule (stop time.txt) of each trip ID. This information can be matched with the AFC and APC-VL data in relation to the time of specific transactions.

Data Preparation

Monday, November 10, 2008, was used as a typical day in the analysis. In detail, the 24-h time span from 3 a.m. on Monday to 3 a.m. on Tuesday was considered because many buses end their trips after midnight, and few trips are overnight. All AFC, APC-VL, and GTFS data were loaded into Microsoft SQL Server 2008. Several conditions were considered in preparing the data. The fare cards (SSNs) had to have at least two transactions on the given day in the AFC data for them to be applied to the trip-chaining model. An inner join was then computed to retrieve all transactions (90,154 for November 10) of each unique SSN because multiple records may have been detected and each SSN may have had a different number of transactions. To clean and validate the retrieved data, several additional filters were applied (discussed in the following section on methodology).

The AFC data itself do not provide either any passenger alighting information or directional information for the route. As a result, the boarding location of the next transaction must be considered to infer the alighting stop of each passenger trip. Although APC-VL data support more accurate identification of boarding stops, these data are only a sample (about 30%), so all the passenger O-D estimates cannot be validated.

Figure 2 shows the kind of data used for this study as well as how they are connected. Each data set can be integrated with one another by using various data relationships to overcome the limitations of each separate data set. The data attribute route number can be used to link the data sets.

AFC data are used to identify the boarding stop. In addition, the nearest stop is found from the APC-VL data for verification purposes. To identify the boarding stop by using APC-VL, the vehicle ID and route number are matched between the AFC and APC-VL data, and the trip number (trip ID) whose scheduled time interval covers the transaction time recorded in the AFC data is found. Finally, distances for all stops having the same trip number are calculated, and the stop with the minimum distance is assigned as the nearest stop.

METHODOLOGY

Assumptions

The approach in this study is mainly based on a trip-chaining model. Therefore, the pertinent assumptions for the trip-chaining model are the most important ones. These, along with some other assumptions made, are discussed below.

Trip-Chaining Model Assumptions

Typical assumptions of trip chaining are made in this study as well as several other studies (*3, 4, 9*). For instance, it is assumed that travelers who use the transit system do not use any other modes within the given sequence of daily transit trips. The major assumption of the trip-chaining model is that the destination of each trip can be inferred from the origin of the next trip. In addition, the destination of the last transaction of a person in a given day is assumed to be the boarding point of that person's first transaction that day.

Once the alighting stop for each transaction is inferred, some proximity checks should also be applied. These checks exclude many of the transactions for which the trip-chaining assumptions are not true. For the geographical check in this paper's algorithm, that the inferred alighting stop was located within walking distance of the next boarding point had to be ensured. For temporal checking, that the inferred alighting time was not later than the next transaction time had to be ensured.

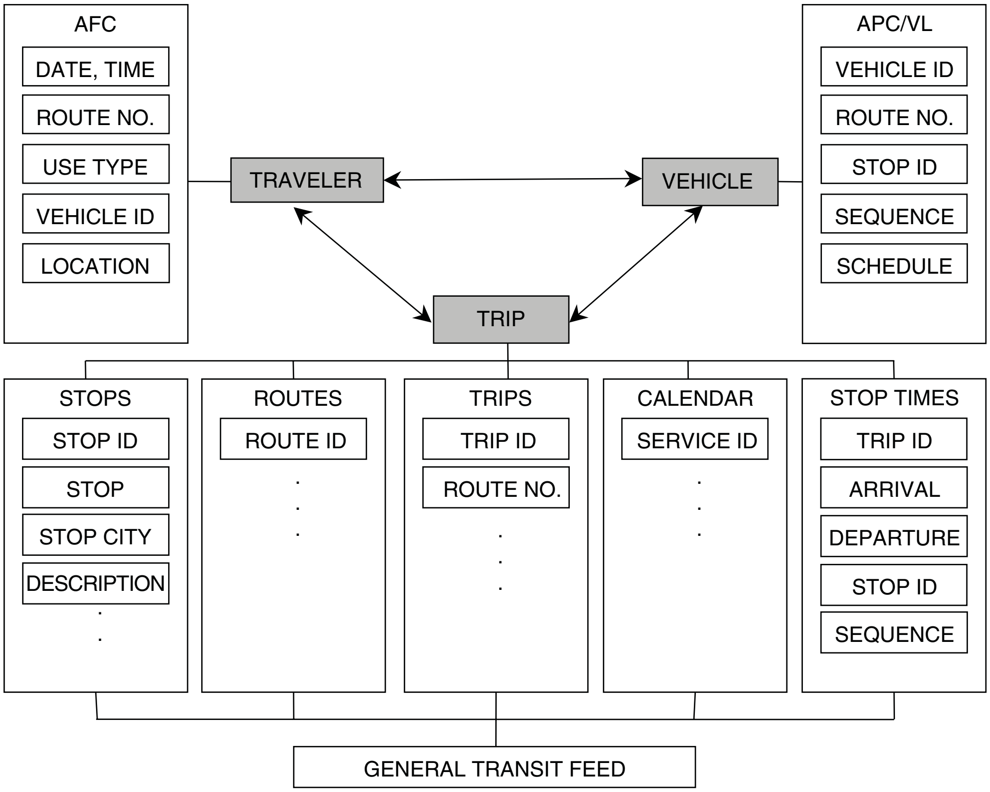

FIGURE 2 Relational schema.

Assumptions of Estimation of Stop-Level O-D

For each person, the origin of the trip is assumed to be the boarding bus stop and the destination is assumed to be the alighting bus stop. It is also assumed that, for the trips that contain transfers, the origin is the boarding stop of the first leg of the trip and the destination is the alighting stop of the last leg of the trip.

Other Assumptions

It is assumed that the required time for an individual to participate in an activity is at least 30 min. It is also assumed that the maximum waiting time for a person to transfer cannot exceed 90 min. Later, this paper explains that the transaction status (transfer or initial trip or use type) of a user is understood by having these criteria considered along with the schedule of the bus route that the user has boarded. For calculation of the walking distance between two successive rides, the Euclidean distance (ED) between the two points was used. To account for nonstraight paths between the two points, ED was multiplied by $\sqrt{2}$, which gave the diagonal of a right-angle metric between the two points. The average walking speed of 3 mph (4.8 km/h) was assumed for estimating the walking time between two points.

Model

Primary Data Refining

The chosen day (Monday, November 10, 2008) began with 90,154 transactions, including both initial and transfer transactions. However, evidence showed that in some cases the AFC transaction data set might have had some wrong or missing entries or might have even been missing one whole transaction in the set of a person's trips. Many of these missing transactions were detected on the basis of transaction status (original use type). For example, if, in the first transaction of a person, the use type was recorded as a transfer, it was inferred that at least one transaction of that person was missing. Eliminated were all transactions of the individual for which it was detected that some transactions might be missing. Of the total 90,154 transactions, 1,970 had such problems, and after all transactions from these fare cards were eliminated, the total number of transactions decreased from 90,154 to 84,413 on the chosen day.

Main Algorithm

The main algorithm proposed for determining passenger O-D stops is shown in Figure 3. Each step in this algorithm is discussed below.

Search Lists

An issue in the algorithm was that a search needed to be done through all the GTFS schedule data (488,105 records) multiple times. Considering the number of transactions (84,413), and to expedite the search process, a search list for each route based on GTFS schedule list was created. After the required GTFS files were combined, the schedule data of the routes were kept in separate lists. Then, for each transaction, it was necessary just to search through the schedule of the specific route to find the boarding and alighting stops.

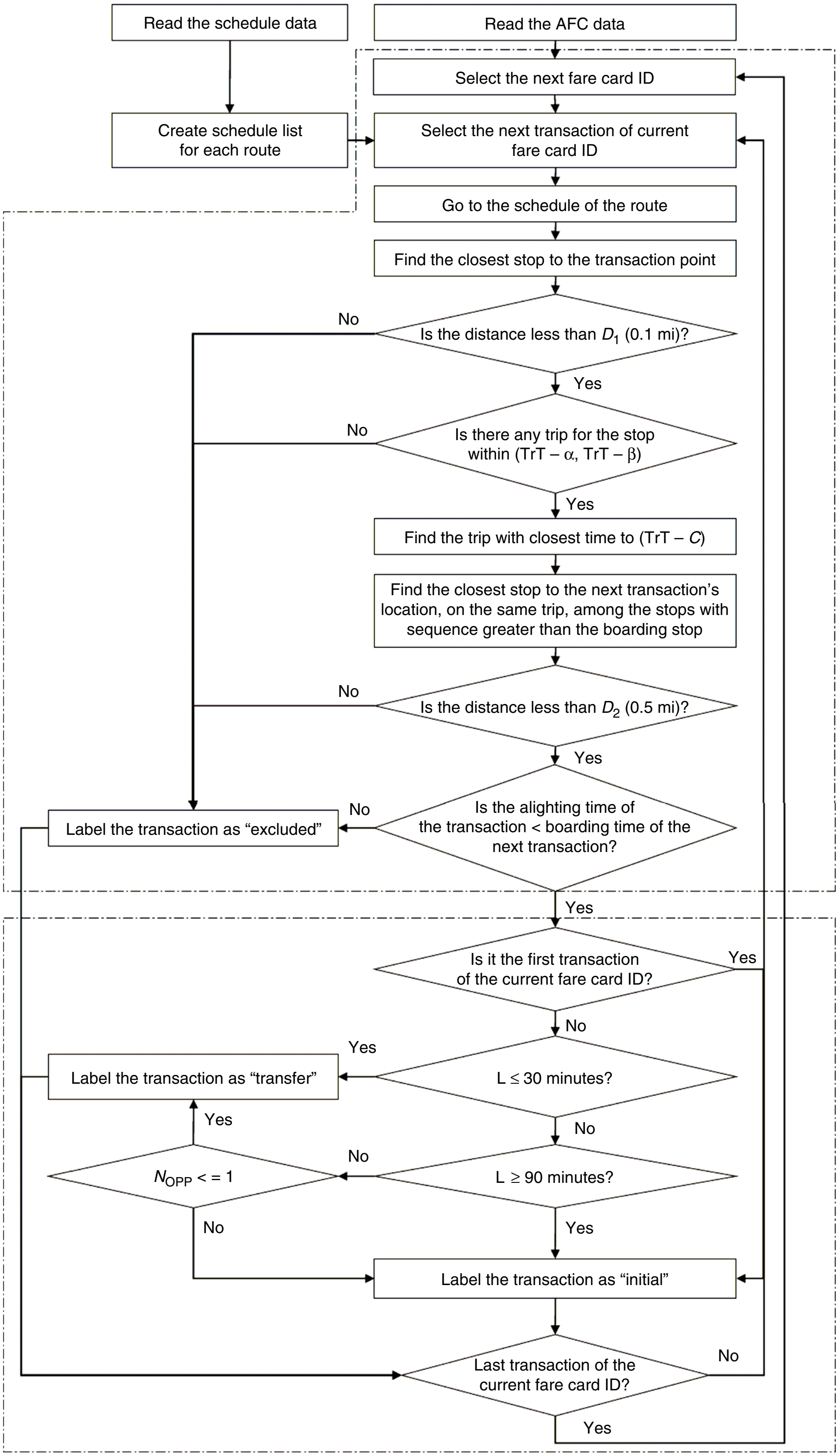
Read the schedule data
Read the AFC data
Select the next fare card ID
Create schedule list for each route
Select the next transaction of current fare card ID
Go to the schedule of the route
Find the closest stop to the transaction point
No
Is the distance less than D_1 (0.1 mi)?
Yes
No
Is there any trip for the stop within (TrT – α, TrT – β)
Yes
Find the trip with closest time to (TrT – C)
Find the closest stop to the next transaction's location, on the same trip, among the stops with sequence greater than the boarding stop
No
Is the distance less than D_2 (0.5 mi)?
Yes
Label the transaction as "excluded"
No
Is the alighting time of the transaction < boarding time of the next transaction?
Yes
Is it the first transaction of the current fare card ID?
Yes
No
Yes
Label the transaction as "transfer"
L ≤ 30 minutes?
Yes
No
No
N_{OPP} < = 1
L ≥ 90 minutes?
No
Yes
Label the transaction as "initial"
Last transaction of the current fare card ID?
No
Yes

FIGURE 3 Main algorithm for estimating transit O-D from AFC data and schedule.

Finding Boarding Stop and Alighting Stop

The first part of the main algorithm (upper dotted box of the Figure 3) is related to finding the boarding stop, trip ID, and alighting stop for each transaction. In this part, for each transaction in the AFC data, the GTFS schedule was searched to find the best-fitting trip ID and boarding and alighting stops.

After the data structure was created, the algorithm found the nearest stop on the specific route to the transaction point and considered this the boarding stop. If the distance between the transaction point and the stop was more than a predefined threshold [due to the nature of Global Positioning System (GPS) accuracy or any other errors], the transaction was labeled excluded. The spatial (or geographical) threshold for checking the boarding stops (D_1) was considered to be 0.1 mi in this study. (In the following section, a sensitivity analysis is performed for D_1 and other parameters.) The next step for the inferred boarding stop was to find the best-fit trip ID. To find it, a statistical analysis was performed to define an appropriate criterion (with Parameter C) by which the most probable trip ID in the schedule could be inferred from the actual time of transaction at the boarding stop.

All transactions were distributed between the actual arrival and departure times of their associated bus run. Then, to find the most probable trip ID in the schedule, the scheduled arrival time from the transaction time at the specific boarding stop was inferred. So an estimate of the average time shift between the transaction time and the scheduled bus arrival time. This average shift (C) is bounded by (a) the average delay between the actual arrival time and the scheduled time and (b) the average delay of the departure from the scheduled time. From the available sample with 18,398 records from APC-VL data, the actual arrival and departure times of buses were found to have an average delay of 26 and 83 s, respectively (Figure 4). Then, a reasonable estimate for C (the time between the scheduled arrival and the transaction time) would be the average of 26 and 83 s, or $C = 54$ s. In the next section, further sensitivity analysis is also performed for Parameter C. With $C = 54$ s, to find the most probable trip ID for the transactions, a search is conducted for the scheduled departure time closest to (TrT – C), where TrT is the transaction time (Figure 5).

In the process of searching for the most probable trip ID, to increase the algorithm speed, a search time interval is considered instead of searching through the whole daily schedule. Under the assumption of a normal distribution for the actual arrival and departure times, the time interval that covers the correct trip ID with a probability of .99 was chosen. This time interval is (TrT – α, TrT – β) where

$$\alpha = \mu_{\text{Dep}} + 2\sigma_{\text{Dep}},$$
$$\beta = \mu_{\text{Arr}} - 2\sigma_{\text{Arr}},$$

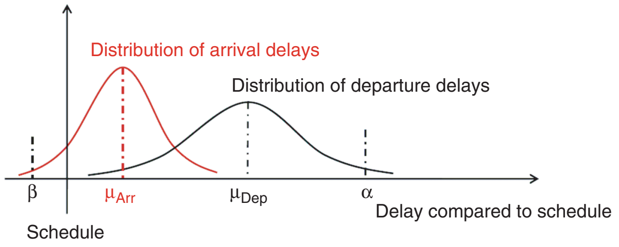

FIGURE 4 Schematic distribution of bus arrival and departure delays and transactions.

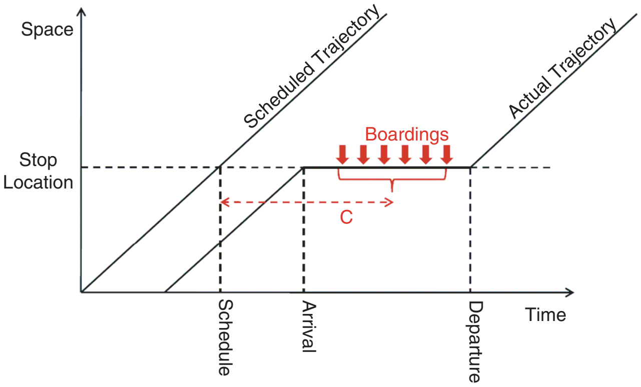

FIGURE 5 Time shift for transactions from schedule, Parameter C.

μ_{Dep} = average delay of bus departures,
μ_{Arr} = average delay of bus arrivals,
σ_{Dep} = standard deviation of delay of bus departures, and
σ_{Arr} = standard deviation of delay of bus arrivals.

The calculated values for α and β from the data set are 383 and –302 s, respectively. Application of these temporal boundaries narrows the search area to the bus runs at the inferred stop and helps improve the accuracy of the inference as well as increasing the search speed. Nonetheless, when GPS accuracy is considered, in some cases it is possible that the wrong stop is selected for the boarding stop, although this possibility cannot be verified with the given data.

Once the best-fit trip ID is found for a boarding stop, the next step is finding the alighting stop for each transaction. To do so, the schedule of the found trip ID is geographically tracked and the stop nearest the next transaction point is located. That stop can be inferred as the alighting stop if the distance between it and the next transaction point is less than a predefined boundary (D_2, which is 0.5 mi in this study). An “excluded” label is placed on the transactions if the suggested alighting stop lies outside this geographical boundary.

Detecting Transfer Trips

Once the boarding stop, alighting stop, and the trip ID are inferred for the transactions in the first part of the main algorithm, the transfer trips among all the transactions have to be detected. The procedure is shown in the dotted rectangle at the bottom of the flowchart in Figure 3.

The original use type attribute for each transaction in the AFC data set specifies whether each transaction is an initial transaction or a transfer, but this specification is not consistent with what is needed to estimate O-Ds. In O-D estimation, the transactions must be grouped in a way that all the transactions in each group form a unique O-D trip (i.e., a so-called linked trip). Under this condition, a unique O-D can be linked to all the transactions in each group. In that case, the first transaction of each group is an Initial one and all the remaining transactions are transfers. But the logic behind the transactions in the Metro Transit data set is not consistent with this purpose. Rather, in the AFC data set, the transactions are grouped into 2.5-h intervals. It can be assumed that this grouping is related to Metro Transit fare policy: because each fare is valid for 2.5 h, once a passenger pays for an initial transaction, he or she can use the system (i.e., make transfers) free of charge for the next 2.5 h. So, in that grouping method, the first transaction of a person is specified as initial and all other transactions that take place within 2.5 h are specified as transfers, regardless of the actual passenger trip purpose.

In this section, to modify the use type in a way to serve the purpose of this study, an attempt is made to scan the spatial and temporal attributes of the travelers in their successive transactions. With consideration of these attributes, an attempt is made to infer the use type for each transaction. Figure 6 shows how the use type for a transaction is modified on the basis of spatial and temporal characteristics of the current transaction and the previous one.

As Figure 6 shows, the geographical and temporal coordination of each transaction in relation to the previous transaction is studied. With the alighting stop for the previous transaction and the boarding stop for the current transaction inferred in the first part of the algorithm, the time–space relationship between the alighting and boarding is now determined. Bus run times are also extracted from the GTFS schedule data, and the departures on each route are considered as time–space points in Figure 6. Then, on the basis of spacing between the previous alighting stop and the next boarding stop, a walking time (W) for the traveler to reach the boarding location is calculated. Also considered is a possible delay (D) due to any setback in alighting or walking or from minor activities like buying a newspaper, coffee, or the like. From a start at the alighting time, and with addition of the walking time and the estimated delay, a time point, t_{acc} (the time from which the boarding stop becomes accessible for the passenger), is inferred. The criteria for understanding use type for the transactions are based on (*a*) the number of bus runs in the time interval from t_{acc} to the actual boarding (transaction) time and (*b*) the time between the estimated arrival time at the boarding stop and the actual boarding time (L).

Similar to the way transfer time criteria were applied by Hofman and O'Mahony (*14*), an upper bound (L_{up}) of 90 min on L for the transfer transaction have been chosen. In addition, under the assumption of a minimum duration of 30 min for an activity, a lower bound (L_{low}) of 30 min on L for the initial transactions has been considered. These criteria mean that one transaction will be interpreted as initial when the calculated L is greater than 90 min and as transfer when the calculated L is less than the minimum expected time for an activity (30 min). When the calculated L for a transaction is between 30 and 90 min, the number of opportunities N_{OPP} available to the passenger for boarding between the estimated arrival time (t_{acc}) and the actual boarding time determines the use type. If $N_{OPP} \leq 1$, the use type is inferred to be transfer; otherwise, it is inferred to be initial. In other words, an $N_{OPP} > 1$ means that the passenger did not board the first accessible bus and implies that an activity has likely occurred before boarding. Through application of these criteria, in the model output with 33,514 transactions, a total of 2,415 transactions previously (in AFC) recorded as transfers were inferred to be initial, and 118 transactions previously recorded as initial were inferred to be transfer.

The combination of the criteria on L with the criterion on N_{OPP} is a major contribution of this study that helps in the consideration of the bus schedule along with the transfer time thresholds in understanding the use type.

Final Refining and Outputs

Once the use type for all the transactions is determined, the transactions of an individual can be divided into different groups. Each group will represent a unique (linked) trip and will have a unique O-D. Each group consists of an initial trip and the dependent transfers (if any). In the output of the algorithm, some transactions exist for which no trip ID, boarding stop, or alighting stop is found. These transactions are labeled excluded in the algorithm. In the final refining step, these transactions are excluded, as are all transactions in the same group with them.

A typical assumption about the AFC system is that the fare is collected from passengers when they board. But for some routes (mostly in express routes originating from the central business district or park-and-ride centers), the fare is collected when passengers alight. Because no information about the bus routes with this characteristic was accessible, suspicious records were eliminated from the output of the model on the basis of the following criteria. These transactions show up in the output with extremely short in-vehicle travel times and with the inferred destinations in the same geographical location as the origins. In the final refining, the transactions using express routes for which the inferred alighting was just one stop from the boarding stop were eliminated.

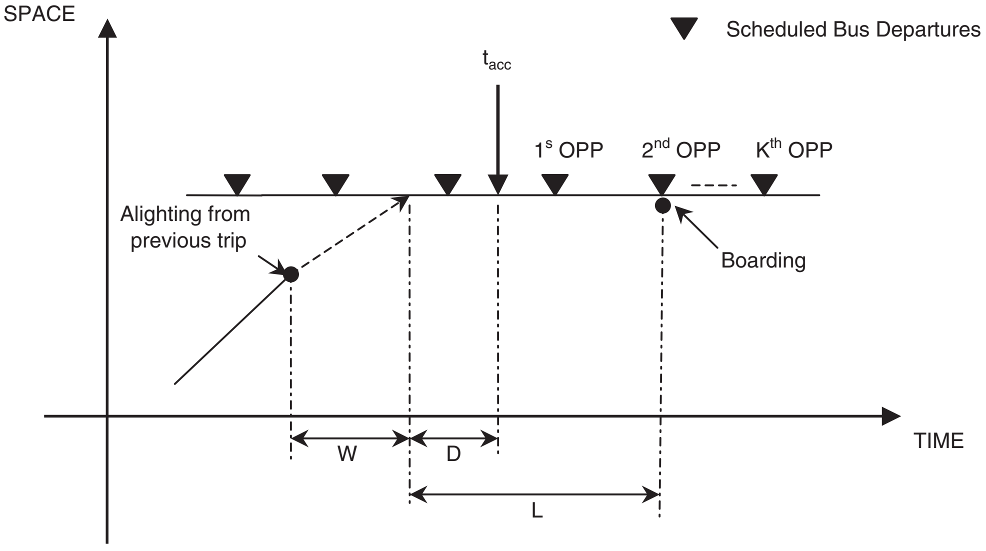

FIGURE 6 Understanding use type.

The total number of the transactions for which the boarding and the alighting stops were found was 51,273, and the total remaining after the final refining was 33,514, which represent 28,260 groups (linked trips). In the output for each of these 28,260 linked trips was an estimated O-D stop pair, which was inferred from the boarding stop on the first leg of the trip and the alighting stop on the last leg of the trip.

Verification of Outputs by Joining AFC with APC-VL Data

After the algorithm was applied and the boarding and alighting stops for each transaction were found, the results were compared with a sample data set consisting of joined AFC–APC-VL data. By using the sample data, it was possible to find the trip that each person had taken, that person's boarding stop, and the correct route direction. In this case, the inferred boarding stop from the algorithm was compared with the boarding stop (if any) found in the merged AFC–APC-VL data, and any possible mismatch could be detected. For the data from Metro Transit, the algorithm resulted in 51,273 transactions, and the rest of them were labeled excluded. The number of records in the matching APC-VL sample was 10,886, which was 21.2% of the available AFC transactions. For 1.6% of the sample data, the direction was inferred incorrectly. These cases were considered the wrong output because the alighting stop would also be inferred incorrectly. The most likely reason for the mismatch was that another mode might have been used between two successive transactions, and the transaction point of the second transaction led to selection of the wrong direction and a wrong alighting point for the first transaction. In 2.9% of sample transactions, although the inferred direction was correct, the inferred boarding stop was not correct (there was no boarding record in the APC-VL data for that stop). The point here is that, for these transactions, the correct direction was selected and it led to the correct alighting stop. For most of these cases (298 of 325 transactions), it was observed that the algorithm's selected stop was the neighboring stop to the correct one, as noted in the APC-VL data. The results of the verification analysis are represented in Table 2. In conclusion, though, the algorithm gave reliable output (correct boarding stop and correct direction) for more than 98% of transactions.

Output Summary

To provide a brief summary of the algorithm's output, the estimated O-D were aggregated for the geographical analysis. Figure 7 presents the O-Ds for the morning (6 to 9 a.m.), midday (9 a.m. to 4 p.m.), and afternoon (4 to 7 p.m.) periods. The origins (morning) and destinations (afternoon) seem to be symmetric, which suggests that fare card holders' trips begin and end at the same locations. During midday, many internal trips within downtown were observed.

TABLE 2 Verification of Model Output in Comparison with APC-VL Sample

Number of Transactions	Percent	Description
10,886	100.0	Matching sample available
10,388	95.4	Verified
325	2.9	Direction is verified, boarding stop is not
298	2.7	Neighboring stop
7	0.2	Not neighboring stop
173	1.6	Direction is not correct

Alternative Submodel for Finding Stops and Trip ID

In the proposed model, once the nearest stop to the location of the transaction is chosen as the boarding stop, the most probable trip ID is taken and the alighting stop is inferred afterwards. But some cases may exist in which, because of the level of GPS accuracy in the AFC data, the nearest stop is not where the transaction has actually happened. Especially when the stop in the opposite direction of the bus route is right across the street from the presumed boarding stop, the GPS may lead to a wrong inference of the stop for boarding in the opposite direction. Such cases in the base algorithm, in the process of distance check, automatically get excluded from the output regardless of whether the boarding stop might be incorrectly inferred.

To manage these cases better and increase the number of trips identified, an alternative algorithm is proposed. If the first algorithm does not output the boarding and alighting stops for a transaction (i.e, the proximity checks do not hold for the inferred stops), before the transaction gets excluded, the alternative algorithm relaxes the search among the stops in the other direction and finds the stop nearest the transaction location. Then, the trip ID for this transaction is chosen and the alighting stop inferred. If the inferred alighting stop is in an acceptable vicinity of the passenger's next transaction, the inferred boarding and alighting stops are confirmed.

After this alternative algorithm was applied to the data set, the total number of inferred transactions increased from 51,273 to 55,714. The difference between the output of the alternative algorithm and the base algorithm is due to consideration of the opposite direction in the procedure for inferring the boarding and alighting stops. This consideration is another contribution in this study.

The main advantages of this alternative algorithm are that it (*a*) increases the size of output, (*b*) detects the cases in which the use of GPS would otherwise lead to an incorrect boarding stop, and (*c*) eliminates any possible bias resulting from the exclusion of these cases from the output.

This model was also applied to the AFC data, and the outputs were generated. However, verification of the outputs, based on their being compared with APC-VL data, was not encouraging versus the base algorithm. The total not-correctly-inferred transactions, for which the direction of the inferred trip ID does not match the direction in the matching sample, increased from 1.6% in the base algorithm to about 4.3% in the alternative algorithm.

SENSITIVITY ANALYSES

Some assumptions were made for the parameters of the model in the previous section. In this section, a sensitivity analysis is done on the parameters for the proposed model. This analysis consists of two parts. First, a change is made in the parameters in the first part of the algorithm including the boundaries for the maximum boarding distance and alighting distance, D_1 and D_2, respectively, and the average time shift, C, between the transactions times and the scheduled arrival times. For these parameters, the results can be compared with the merged AFC–APC-VL sample data to decide which values give better results. Second, the second part of the algorithm that deals with the use type inference is analyzed. For parameters of this part,

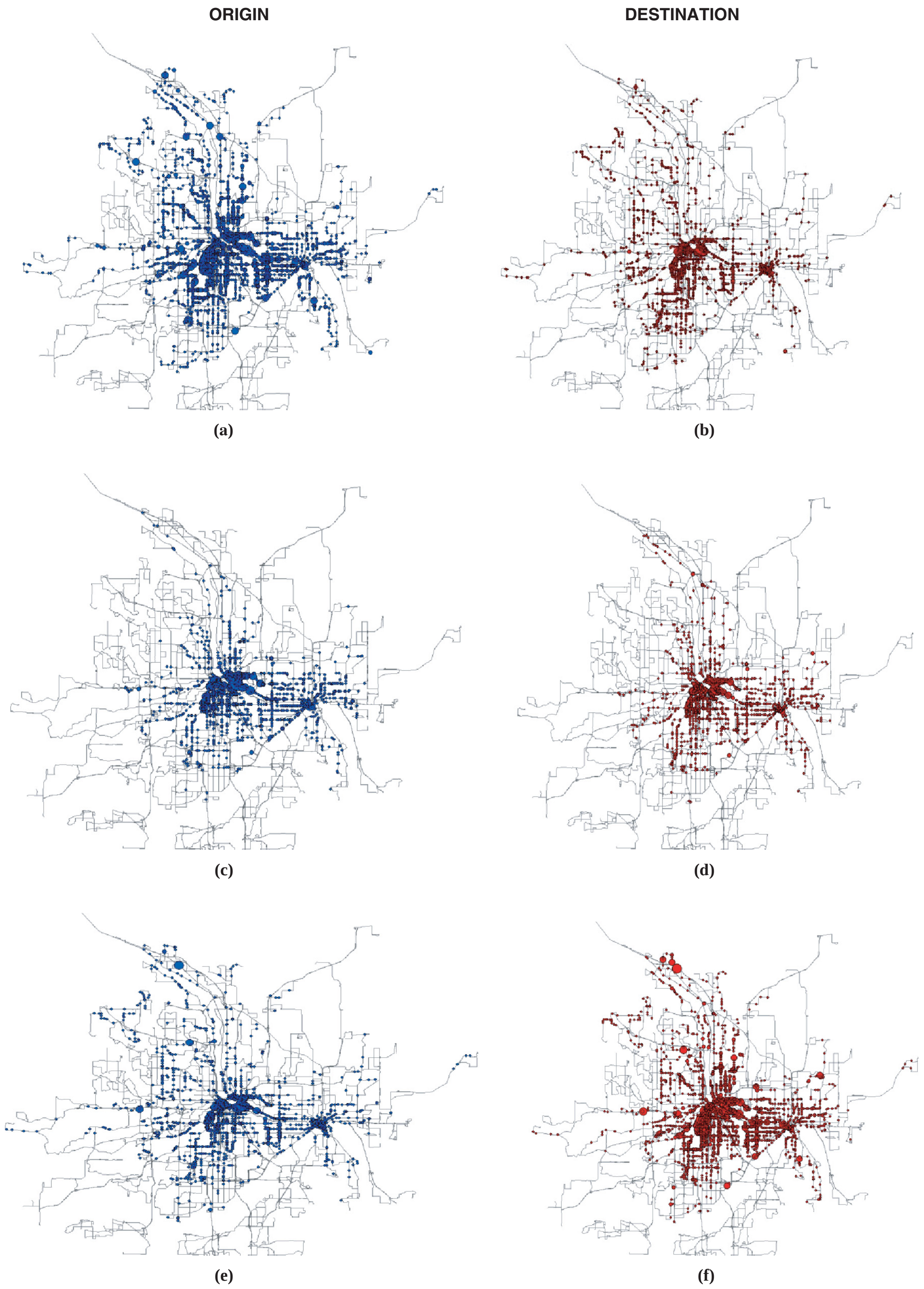

FIGURE 7 Geographical analysis of (*a*) morning origins, (*b*) morning destinations, (*c*) midday origins, (*d*) midday destinations, (*e*) afternoon origins, and (*f*) afternoon destinations.

TABLE 3 Sensitivity Analysis on Parameters in First Part of Proposed Model

	C (s)								
	39			54			69		
D_1 (miles)	0.25	0.1	0.1	0.25	0.1	0.1	0.25	0.1	0.1
D_2 (miles)	0.5	0.5	0.25	0.5	0.5	0.25	0.5	0.5	0.25
Number of transactions with inferred boarding and alighting stops	51,940	51,264	46,417	51,946	51,273	46,410	51,943	51,269	46,398
Percentage of transactions with wrong inferred direction	1.7	1.6	1.1	1.7	1.6	1.1	1.7	1.7	1.2

including the lower bound of the initial trip waiting time, L_{low}, the upper bound of transfer waiting time, L_{up}, and the possible access delay, D, there is no source for verification. So the objective of the analysis for these parameters is to see whether the proposed model is sensitive to these parameters.

Different values for parameters of the first part of the algorithm, including D_1, D_2, and C, are chosen, and the model is run for each combination. Then, a verification analysis (like that shown in the section on methodology) is done, and the percentage of unacceptable results is calculated (see Table 3). Results show that the model is not sensitive to Parameter C, and the output does not change significantly with slight changes in this parameter. This insensitivity is mostly due to the large headways in the transit system. But changing Parameters D_1 and D_2 slightly affects the outputs. The best results are gained by using 0.1 and 0.25 mi for D_1 and D_2, respectively, on the basis of the number of transactions with an incorrectly inferred direction. But these values decrease the total number of accepted transactions by 5%. In conclusion, because the model is not extremely sensitive to these parameters, the chosen values for D_1 and D_2 (0.1 and 0.5 mi, respectively) seem to be reasonable.

The second part of the sensitivity analysis is on the parameters affecting the use type inference, including L_{low}, L_{up}, and D. Different values were chosen for the parameters, and the percentage of transactions inferred as initial over all remaining transactions, after the final refinement was applied, was calculated. Investigated were values of (*a*) L_{low} from 20 to 40 min in 10-min increments, (*b*) L_{up} from 60 to 120 min in 30-min increments, and (*c*) D from 0 to 20 min in 5-min increments. Results showed that the model was not sensitive to these parameters because, over all the combinations, the percentage of transactions inferred as initial ranged from 84.5% to 86.2%.

CONCLUSIONS AND FUTURE WORK

By using AFC, GTFS, and APC-VL, a model was created to infer boarding and alighting stops as the route direction was considered. Application of the model to the data set found appropriate boarding and alighting stops for 51,273 of 84,413 transactions (gleaned from an initial 90,154). Then by application of some criteria to detect transfers, the use type of each transaction was understood. In the final refining process, the final output size decreased to 33,514 transactions, which belong to 28,260 (linked) trips. By comparison with an AFC–APC-VL matching sample of 10,886 transactions, the output of the main model was verified in more than 98% of transactions.

An alternative model was also established during this study, and it improved the algorithm by increasing the number of outputs; the improvement resulted from consideration of both directions of each route. But on the basis of the accuracy of the inference, the main model was preferred. Although at issue is exclusion of the transactions that have been guessed to be incorrectly inferred by the proposed model, which decreases the output size, a trade-off exists between the output size (quantity) and its accuracy (quality). The choice depends on the researcher's perspective of how to approach this issue because (*a*) it is difficult to capture all travelers' behaviors accurately and (*b*) the data may have some inconsistencies with the trip-chaining assumptions. Finally, a sensitivity analysis was performed for the parameters used in both parts (finding the stops and understanding the use type) of the model, and this analysis showed that the model is not sensitive to any of the parameters.

The outcome from this research can lead to related work. First, the O-D estimation can be extended to include accessibility by using walking distance–time to the boarding and alighting stops. This stop-level O-D estimation should be expanded to a zone- or parcel-level O-D estimation because the activities do not originate from a stop but from home or attraction points. Second, as the stop-level O-D and its possible paths between O-D stops are secured, it is possible to set up a utility model and empirically estimate a path choice model. Third, threshold estimation, especially for transfers, is another promising future research area that considers trip length–travel time distribution. In the main model, 30- and 90-min thresholds were applied for detecting transfer behavior. It should be possible to adjust the static boundary in accordance with network configuration and transit passenger behavior. Fourth, a comparison with other cities that use the smart card system is another area for additional research. As more ADC systems are put into service globally, comparisons of different ones can provide interesting work.

ACKNOWLEDGMENTS

The data in this study were graciously provided by Metro Transit and assembled by Chen-Fu Liao of the University of Minnesota and Nigel Wilson of the Massachusetts Institute of Technology. The authors are particularly indebted to them for their work in obtaining these data. In addition, the authors are grateful to Paul Krech and John Levin of Metro Transit's scheduling department, who generated a GTFS data set from the November 2008 schedules.

REFERENCES

1. Cui, A. Bus Passenger Origin–Destination Matrix Estimation Using Automated Data Collection Systems. MS thesis. Massachusetts Institute of Technology, Cambridge, 2006.
2. Farzin, J. M. Constructing an Automated Bus Origin–Destination Matrix Using Farecard and Global Positioning System Data in São Paulo, Brazil. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2072,* Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 30–37.

3. Seaborn, C., J. Attanucci, and N. H. M. Wilson. Analyzing Multimodal Public Transport Journeys in London with Smart Card Fare Payment Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2121*, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 55–62.
4. Zhao, J. *The Planning and Analysis Implications of Automated Data Collection Systems: Rail Transit OD Matrix Inference and Path Choice Modeling Examples.* MS thesis. Massachusetts Institute of Technology, Cambridge, 2004.
5. Zhao, J., A. Rahbee, and N. Wilson. Estimating a Rail Passenger Trip Origin–Destination Matrix Using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering,* Vol. 22, 2007, pp. 376–387.
6. Reddy, A., A. Lu, S. Kumar, V. Bashmakov, and S. Rudenko. Entry-Only Automated Fare Collection System Data Used to Infer Ridership, Rider Destinations, Unlinked Trips, and Passenger Miles. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2110,* Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 128–136.
7. Chapleau, R., M. Trépanier, and K. K. Chu. The Ultimate Survey for Transit Planning: Complete Information with Smart Card Data and GIS. Presented at 8th International Conference on Survey in Transport, Lac d'Annecy, France, 2008.
8. Chu, K. K. A., and R. Chapleau. Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2063,* Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 63–72.
9. Chu, K. K. A., R. Chapleau, and M. Trépanier. Driver-Assisted Bus Interview: Passive Transit Travel Survey with Smart Card Automatic Fare Collection System and Applications. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2105,* Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 1–10.
10. Trépanier, M., N. Tranchant, and R. Chapleau. Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems: Technology, Planning and Operations*, Vol. 11, No. 1, 2007, pp. 1–14.
11. Bagchi, M., and P. R. White. The Potential of Public Transport Smart Card Data. *Transport Policy,* Vol. 12, 2005, pp. 464–474.
12. Google. *General Transit Feed Specification.* http://code.google.com/transit/spec/transit_feed_specification.html. Accessed June 2010.
13. *GTFS Data Exchange*. www.gtfs-data-exchange.com. Accessed June 2010.
14. Hofmann, M., and M. O'Mahony. Transfer Journey Identification and Analyses from Electronic Fare Collection Data. *Proc. 8th International IEEE Conference on Intelligent Transportation Systems*, Vienna, Austria, Sept. 2005, pp. 13–16.

The Transportation Network Modeling Committee peer-reviewed this paper.