A robust method for estimating transit passenger trajectories using automated data

Pramesh Kumar[a], Alireza Khani*[b], Qing He[c]

[a]*Department of Civil, Environmental and Geo-Engineering, University of Minnesota, Twin Cities*
[b]*Department of Civil, Environmental and Geo-Engineering, University of Minnesota, Twin Cities*
[c]*Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York*

Abstract

Development of an origin-destination demand matrix is crucial for transit planning. The development process is facilitated by automated transit smart card data, making it possible to mine boarding and alighting patterns on an individual basis. This research proposes a novel trip chaining method which uses Automatic Fare Collection (AFC) and General Transit Feed Specifications (GTFS) data to infer the most likely trajectory of individual transit passengers. The method relaxes the assumptions on various parameters used in the existing trip chaining algorithms such as transfer walking distance threshold, buffer distance for selecting the boarding location, time window for selecting the vehicle trip, etc. The method also resolves issues related to errors in GPS location recorded by AFC systems or selection of incorrect sub-route from GTFS data. The proposed trip chaining method generates a set of candidate trajectories for each AFC tag to reach the next tag, calculates the probability of each trajectory, and selects the most likely trajectory to infer the boarding and alighting stops. The method is applied to transit data from the Twin Cities, MN, which has an open transit system where passengers tap smart cards only once when boarding (or when alighting on pay-exit buses). Based on the consecutive tags of the passenger, the proposed algorithm is also modified for pay-exit cases. The method is compared to previous methods developed by the researchers and shows improvement in the number of inferred cases. Finally, results are visualized to understand the route ridership and geographical pattern of trips.

Keywords: Automatic Fare Collection (AFC), General Feed Transit Specification (GTFS), Transit Origin-Destination (O-D) Matrix, Transit, Trip Chaining Algorithm, Smart Card Data

1. Introduction

For better service and planning, transit agencies need to understand passengers' travel behavior. For this purpose, they conduct on-board surveys which collect data about passengers' boarding and alighting location, purpose of travel, etc., and then use expansion factors to expand the survey data for the whole population. There are various limitations associated with these surveys, such as cost, small sample size, bias, and other general reporting errors (Attanucci, J. & Wilson 1981). Conversely, automated data collection systems (ADCS), which are designed for administrative purposes such as revenue management, provide a rich source of information about passengers travel pattern on an individual basis. The automated data offers several advantages (Wang et al. 2011) over traditional surveys by:

1. providing a link to passenger's trips over a longer period of time
2. providing information about the share of different transit commuters (e.g. students, workers, etc.)
3. storing the information in SQL database systems and using it efficiently
4. providing various research opportunities for analyzing passengers' travel pattern

In recent years, there has been growing interest in using automated smart card data for travel behavior research in transit systems. Automatic Fare Collection (AFC) systems collect information about on-board transaction of passengers such as boarding stop/station, date and time of the transaction, route information, etc. The data is useful not only for improving day-to-day transit operations but also for long-term strategic planning of transit network (Pelletier et al. 2011). It has been used for a variety of purposes such as:

1. stop-level origin-destination matrix estimation (Barry et al. 2007, Trépanier et al. 2007, Zhao et al. 2007, Alfred Chu and Chapleau 2008, Barry et al. 2009, Chu and Chapleau 2010, Wang et al. 2011, Nassir et al. 2011, Munizaga and Palma 2012, Gordon et al. 2013).
2. trip purpose inference (Lee and Hickman 2014, Kusakabe and Asakura 2014, Alsger et al. 2018)
3. route choice modeling (Kim et al. 2017, Zhao et al. 2017)
4. passenger trip prediction (Zhao et al. 2018)
5. mining spatial and temporal clusters of similar travel patterns (Ma et al. 2013, Briand et al. 2017, Khani 2018)
6. passenger waiting time estimation (Ingvardson et al. 2018)

This study focuses on one of the important input for analyzing a public transit system, which is the flow of passengers between different stations/stops known as an origin-destination (O-D) matrix. O-D estimation using automated smart card data has attracted attention of many researchers over the last decade (Barry et al. 2007, Trépanier et al. 2007, Zhao et al. 2007, Alfred Chu and Chapleau 2008, Farzin 2008, Barry et al. 2009, Chu and Chapleau 2010, Nassir et al. 2011, Wang et al. 2011, Ma et al. 2012, Munizaga and Palma 2012, Gordon et al. 2013, He and Trépanier 2015). The estimation requires a sequence of trips made by the passenger throughout the day recorded using AFC system. But the information available with this data is limited and the full sequence of trips is usually not available. This is because of the type of the fare collection system (open or closed) employed

by a transit agency. In closed transit systems (Alsger et al. 2016), origin and destination is known for the trips as passengers tap their card both when boarding as well as when alighting, whereas in open transit systems (Barry et al. 2007, Trépanier et al. 2007, Zhao et al. 2007, Alfred Chu and Chapleau 2008, Barry et al. 2009, Chu and Chapleau 2010, Nassir et al. 2011, Wang et al. 2011, Munizaga and Palma 2012, Gordon et al. 2013), the boarding of passengers is usually known, and the alighting is unknown as passengers only tap their card when boarding a transit vehicle. Passengers' alighting location can then be inferred based on the next boarding location using a trip chaining algorithm (Barry et al. 2007, Trépanier et al. 2007, Zhao et al. 2007, Alfred Chu and Chapleau 2008, Farzin 2008, Barry et al. 2009, Chu and Chapleau 2010, Nassir et al. 2011, Wang et al. 2011, Munizaga and Palma 2012, Ma et al. 2012, Gordon et al. 2013, He and Trépanier 2015, Kumar et al. 2018).

Trip chaining algorithms developed so far use assumptions on various parameters, e.g. buffer radius to find the closest stop to the boarding location, walking distance threshold after alighting to board the next route, time threshold to distinguish between boarding and transfer, etc. These parameters can vary among different transit systems and can affect the trip chaining results and therefore the origin-destination matrix. The current research tries to relax the assumptions related to these parameters by proposing a robust trip chaining algorithm.

The algorithm is applied to the AFC data from Twin Cities, Minnesota which has an open transit system (Nassir et al. 2011), where transit passengers use (tap) their card only once. The system is more complex than other systems described in previous research because sometimes passengers tap their card while entering the bus (when they board a "regular route" or "non pay-exit" bus) or sometimes while exiting the bus (when they alight a "pay-exit" bus). The pay exit buses are generally outbound trips from central areas such as Downtown Minneapolis or the University of Minnesota campus to sub-urban areas. The existing trip chaining algorithm changes significantly when the combination of such tags are observed for a card number. The proposed method creates a set of possible trips for a given card tag, calculates the probability that the passenger has used each trip, and then infers the boarding and alighting on the basis of the most likely trip.

The rest of the paper is organized as follows: Section 2 presents a summary of related work done in this research area, followed by motivation behind this research in Section 3. Then, the proposed trip chaining algorithm is described in Section 4, which is followed by the analysis of the results in Section 5. Finally, conclusions and recommendations for future research are provided in Section 6.

2. Related Work

As most of the fare collection systems record passengers' boarding information only, alighting information must be inferred using the sequence of taps (or tags) made by the passenger throughout the day. Thus, a significant amount of research has been done to develop algorithms to determine the alighting location (Li et al. 2018). Navick and Furth

2002 used location-stamped fare box data of Los Angeles area bus routes to determine alighting location using an assumption that boarding pattern of current trip and alighting pattern of opposite trip are symmetric for the entire day which means passengers board the bus again from the same stop where they alighted during the previous trip. Building on that assumption, Zhao et al. 2007, Barry et al. 2007, Barry et al. 2009, and Gordon et al. 2013 developed a method of trip chaining for origin and destination inference with the following assumptions:

1. passengers return to the same location to board the bus where they alighted during the previous trip,
2. no private mode of transportation is used between trips,
3. passengers do not walk a long (more than a certain threshold) distance to board a bus or train,
4. passengers end their last trip at the same location where they started their journey of the day.

Based on the above assumptions, Trépanier et al. 2007 proposed a model which infers alighting stops by minimizing the distance between the alighting stop of the current trip and boarding of the next trip. They applied their method on AFC data from Quebec, Canada and inferred 66% of the trips. Similarly, Wang et al. 2011 proposed a method which combines Automatic Vehicle Location (AVL) data with AFC data from London to infer the origin and destination of different trips and validated the results using bus passenger origin and destination survey (BODS) data. Then Seaborn et al. 2009 stated some rules for trip chaining such as maximum acceptable transfer time of 20 minutes for underground subway-to-bus, 35 minutes for bus-to-underground subway, and 45 minutes for bus to bus trips. Building on the work of Seaborn et al. 2009 and Wang et al. 2011 in estimating origin-destination matrix using London smart card (Oyster) data and iBus vehicle location data, Gordon et al. 2013 specified the importance of the return trips, bus wait time, repeated service and circuity in trips. The researchers suggested a circuity rule to account for the return trips. By using 750m as the maximum alighting distance, circuity factor of 1.7 and minimum transfer time of 5 minutes and maximum time from 30 to 90 minutes, they inferred 96% of the boarding locations and 74.5% of the alighting locations.

Nassir et al. 2011 used AFC data with General Transit Feed Specification (GTFS) data (Google 2005) instead of commonly used AVL data to infer origins and destinations. They used the closest stop found within an upper bound distance of the smart card tag location as the boarding. Using the route information given in the AFC tag (transaction), a search is done for a trip closest in time within an interval of AFC transaction time. Using that trip, the stop found closest to the next boarding is inferred as the alighting stop given that the distance between inferred alighting and next boarding is less than 0.5 miles. Gordon et al. 2018 extended the research on origin-destination estimation of smart card users to non-smart card transit users. They proposed a scaling method for expanding the OD matrix using the fare box data from London and compared the results with the Iterative Proportional Fitting (IPF) method. Luo et al. 2017 and Ma et al. 2013 used the AFC data to produce an aggregate O-D matrix.

Researchers have also tried to validate the trip chaining assumptions either by doing a survey (Seaborn et al. 2009, Wang et al. 2011) or using data from closed transit systems (where passengers tap their card both when entering and as well as exiting the station) (Alsger et al. 2016). For example, Farzin 2008 validated the assumptions of the closest stops and daily symmetry using a travel diary survey in New York, which showed 90% accuracy. Similarly, Alsger et al. 2016 used South-East Queensland public transport smart card data, which has both boarding and alighting information, to implement and validate the current trip chaining algorithms. The researchers also suggested some improvements in the current algorithm, e.g. the alighting of the last tag on a day is the stop nearest to the first boarding of the day on the given transit route. They also suggested the average distance between the actual and estimated alighting stops as 0.33 miles instead of 0.5 miles. Of course, this distance parameter can vary for different transit systems, which we try to relax in this study.

Recent research on trip chaining has pointed out some limitations in trip chaining algorithms and suggested some improvements. For example, Munizaga and Palma 2012 identified that wrong alighting can be inferred if a passenger takes a bus which runs in both directions to go a few blocks away because the passenger would just cross the street to board the next bus rather than taking a long route in the opposite direction. To alleviate this problem, the researchers suggested a cost function which is the sum of the current transaction time and the walking time multiplied by some penalty factor obtained from a discrete choice model. The adopted methodology inferred 80% of the trips using data from Santiago, Chile. The algorithm proposed in the current paper avoids such situations by discarding the trip which is less likely to be taken by the passenger. He and Trepaniér followed their previous work, Trépanier et al. 2007, and proposed a method to infer the boarding and alighting of unlinked trips. The method multiplies the temporal and spatial probabilities calculated using historical location and time of tags to infer the potential alighting.

The quality of trip chaining results depends on fare collection system correctly recording the tag information which is assumed to be correct by most of the studies. This assumption may result in wrong inference of boarding, alighting or especially transfer detections. Robinson et al. 2014 pointed out various causes for why different systems may not record correct information. The possible causes are AVL system failure, card reader failure, software failure, etc. They proposed a method to identify such erroneous smart card data and suggested where transit agencies should target resources to enhance the performance of their AVL and AFC systems. They applied the proposed method to Singapore smart card data and found that alighting for about 7.7% of the tags was found one stop before the actual alighting location and for 0.7% of the tags, the alighting location was found one stop after the actual alighting.

While applying the current trip chaining algorithms to the Twin Cities' AFC data, similar errors in results were found. To improve the accuracy of the results, the current research proposes a robust trip chaining method to alleviate the effect of various assumptions on the parameters such as GPS inaccuracy (buffer zone for boarding stop inference), finding most likely trip from GTFS data, etc. The method is similar to the one used for map matching problem for multi-modal transportation network modeling (Perrine et al. 2015) and can be

applied to other transit systems with any smart card data structure. The research also deals with complex transit systems consisting of "pay-exit" buses (passengers tap their card while alighting) in the Twin Cities, in which case passengers' alighting is known but not their boarding.

3. Motivation

This section explains the motivation behind this research, i.e. the problems and the desired improvements in a current trip chaining algorithm developed by Nassir et al. 2011. The algorithm uses GTFS data (Google 2005) instead of AVL data because the currently available AVL data for the Twin Cities transit system gives the vehicle location on time point stops only instead of all stop locations along a route. Widespread use of GTFS is one of its advantages, making it more readily available than AVL data. Schedule adherence information from AVL data is also used to supplement the GTFS data. Note that the algorithm uses consecutive tags of a card holder which are termed as "current" and "next" tag throughout this paper. For the last tag of the day, next tag can be assumed as the first tag of the day. First, the trip chaining algorithm developed by Nassir et al. 2011 is summarized below:

1. Read AFC data and select the current and next tags.
2. Extract GTFS schedule of the current tag's route and direction to find the closest stop to the current tag location.
3. Go to step 4 if the distance between the current tag and closest stop found is less than 0.1 miles otherwise exclude the tag and go back to step 1.
4. Find a trip within TrT–α and $TrT + \beta$ closest to the current tag time. Here, TrT is the current tag time and α and β are schedule adherence parameters determined using Automatic Passenger Count-Vehicle Location (APC-VL) data.
5. Find the closest stop to the next tag location on the trip found in step 4 for the stops sequence greater than the stop found in step 2.
6. Go to step 7 if the distance between the inferred alighting location of the current tag and the next tag location is less than 0.5 miles, otherwise exclude the tag.
7. Go to step 8 if the boarding time of the next tag is greater than the alighting time of the previous tag, otherwise exclude the tag and go to step 2.
8. Determine if the current tag is the first tag of the day. If it is, mark it as "boarding", otherwise determine if it is a transfer. A detailed discussion about transfer detection is given later in this paper.

The method, although working in most of the cases, may result in wrong inference or no inference in some cases. These cases are described below.

3.1. The sub-route problem

To manage some of the transit routes efficiently, the Twin Cities transit system has sub-routes for most of the high frequency routes. For example, route 2 has sub-routes 2A, 2C, 2E and route 3 has sub-routes 3A, 3B, 3C, 3E, 3K. Generally, one of the sub-routes is more common than the others and runs throughout the day, whereas others are either short turns

or branches to serve more areas. To better understand the sub-route problem, let us consider an instance (Figure 1).

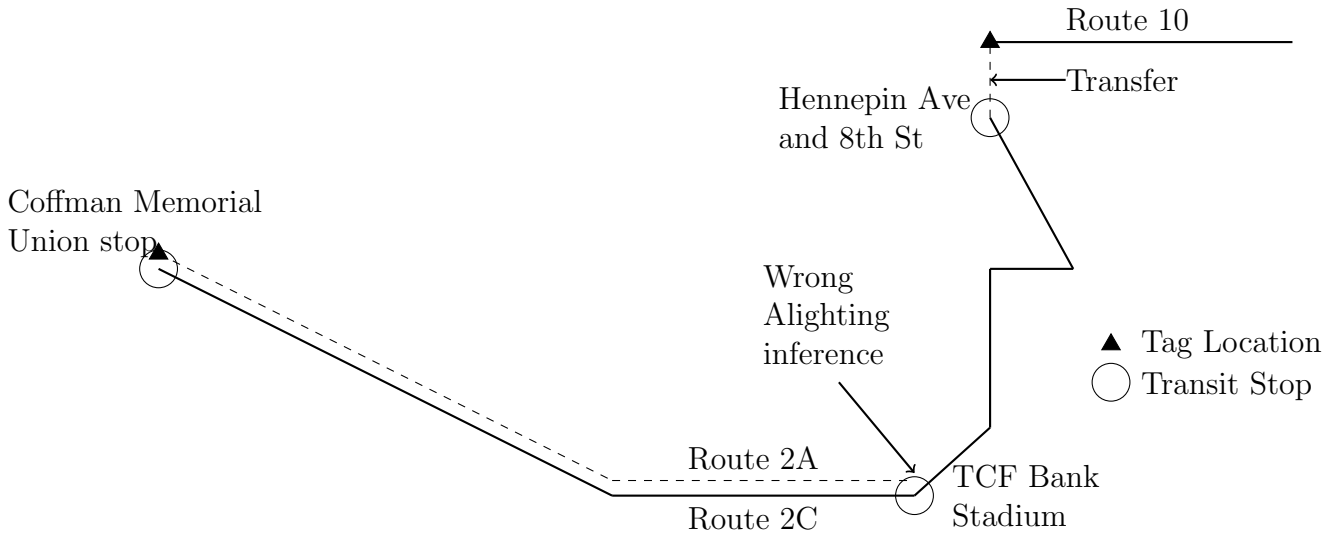


Figure 1: Incorrect alighting inference due to selection of incorrect sub route

A passenger took the bus route 2 from Coffman Memorial Union stop and alighted at Hennepin Ave and 8th Street to transfer to route 10. The current trip chaining algorithm selects any trip from GTFS data which is closest in time to the current tag time. If it selects the trip within route 2A that only goes up to TCF Bank Stadium stop and infer it as alighting stop, then the distance between this stop and the next tag location is more than the walking distance threshold and the algorithm does not infer any alighting stop (discards this record). In this case, a more robust inference method is required to correctly infer the trip within route 2C, which connects with route 10 at Hennepin Ave and 8th St.

3.2. The boarding stop inference problem

The GPS location of tags provided by AFC system may consist of location measurement errors (Robinson et al. 2014). If the algorithm simply finds the closest stop to the tag location, then a potentially wrong boarding stop inference may result in wrong trip inference, wrong alighting stop inference or no inference at all.

3.3. The "pay-exit" problem

Because of high commuter demand to Downtown Minneapolis, Downtown St. Paul, and the University of Minnesota campus, some of the outbound bus routes in the evening peak let passengers enter the bus while boarding and pay while alighting (unlike the regular routes where riders tap while entering the bus). Such cases were not considered during previous studies. In these cases, we do not know the boarding but know the alighting location. Depending on the combination of tags made by a passenger throughout the day, missing boarding or alighting may or may not be inferred. This arises four different cases depending on the consecutive tags of the passenger (Figure 2).

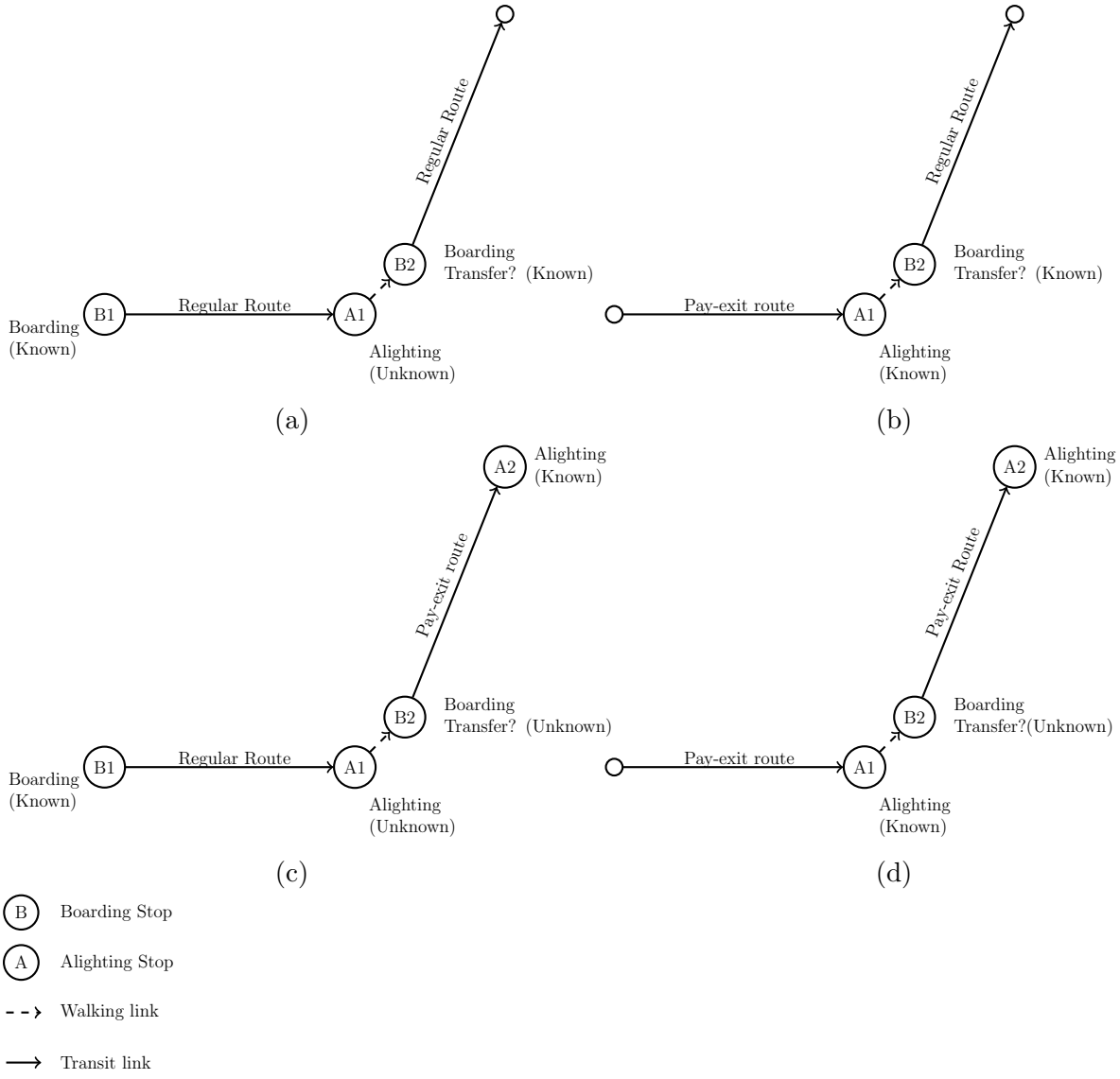


Figure 2: Four cases depending on the pay exit or regular route

1. Current tag (B1) is regular and next tag (B2) is regular
   This is the normal case which has been considered previously in the research. Here, we know the boarding of the current as well as the next tag. Using the route and direction information of the current tag, we can infer the alighting location of the current tag.
2. Current tag (A1) is pay exit and next tag (B2) is regular
   In this case, we know the alighting of the current tag and boarding of the next tag. This is the easiest case among four cases as we need not to infer any location. The only thing to determine in this case is to detect whether or not the next tag is a transfer. Note that the possibility of inferring the boarding of the current tag depends on its previous tag. Similarly, the possibility of inferring the alighting of the next tag depends on its next tag.
3. Current tag is regular (B1) and next tag (A2) is pay exit
   This is the most difficult case among all as we know the boarding of the current tag and the alighting of the next tag which means alighting of the current tag and the boarding of the next tag is missing. Two sub-cases arise in this case depending on the bus route used.
   - If two different bus routes (which are not geographically parallel) are used for both tags, then we can find stops connecting two routes which gives the least distance between the inferred alighting of the current tag and the inferred boarding of the next tag.
   - If same or parallel routes are used for both tags, then we cannot infer the alighting of the current tag and boarding of the next tag. This sub case is quite usual for commuters who take a bus from sub-urban areas which is regular in the inbound direction in the morning but when they return to their home, the same bus is pay exit in the outbound direction in the evening. We propose a method of proportion later in this paper to approximate these cases.
4. Current tag is pay exit (A1) and next tag (A2) is pay exit
   In this case, we know the alighting of both current and next tag. We can make a search list of the stops that come before the alighting stop of the next tag and infer the boarding of the next tag by finding the stop closest to the alighting location of the current tag. Again, the boarding of the first tag may or may not be inferred depending on its previous tag.

4. The Robust Trip Chaining Algorithm

The proposed method for trip chaining in this paper is similar to map matching algorithms used for multi-modal transportation network modeling (Li 2012, Perrine et al. 2015). The map matching algorithm is used to map the public transit stops from GTFS data to a road network by creating a restricted shortest path problem. In this way, it avoids the problems like complicated road geometry, and lack of dynamic vehicle information like vehicle trajectory, speed, turning and heading. Similar methods are common for matching GPS locations to existing road networks to track the trajectory of a vehicle using probability models such as Hidden Markov Model (Newson and Krumm 2009). The proposed trip chaining method also finds a set of candidate trips for a given AFC tag to reach the

next tag, calculates the probability of each trip, then the most likely trip is found to infer the boarding and alighting stops. In this way, different problems faced by the current trip chaining algorithm are addressed. We start with the basic case when both of the consecutive tags are regular which can be applied to any transit system and then we can expand this method to specific cases for the Twin Cities data.

Table 1: Notations used in the paper

Variable	Definition
n	Index/row number in the AFC data
t	Time of the tag
r	Bus route number of the tag
δ	Direction of bus route
θ	Geographical coordinates of the tag
$\mathcal{GC}$	Great circle distance
α	Buffer distance for finding possible boarding stops
ϵ	Buffer distance for finding possible alighting stops
τ	Buffer time for finding possible trips
k	Index for different boarding stops
l	Index for different trips
m	Index for different alighting stops
S_n	List of possible boarding stops for tag n
$\mathcal{T}_{nk}$	List of possible trips for tag n and boarding stop k
Δ_{kl}	Absolute difference between tag time t_n and trip time $t_{tr_{kl}}$
A_{nkl}	List of possible alighting stops for tag n, boarding stop k and trip l
$\mathcal{IV}_{klm}$	In-vehicle travel time for trip l with boarding stop k and alighting stop m
w_{klm}	Walking distance from alighting stop m for trip l with boarding stop k to the next tag location θ_{n+1}

4.1. Trip set generation

Consider two consecutive tags n and $n+1$ of a particular card number on a given date. Using GTFS data, we can make a list of candidate stops $S_n = \{s_{nk}, k = 1, 2, ...\}$ found within a buffer distance of α miles of the tag location θ_n given route r_n and direction δ_n. The value of α can be suitably taken depending on the accuracy of the GPS. For example, previous studies have used $\alpha = 0.1$ miles to find the boarding stop. This will consider the possibility of all the stops which are close to the tag location θ_n being the boarding stop and help in obviating the problem of wrong boarding stop being selected. The error in the GPS location is usually modeled using great circle distance (Newson and Krumm 2009) which is the shortest distance between two points on the surface of a sphere (Navy 2008). We can find the great circle distance d_{nk} between θ_n and s_{nk} as

$$d_{nk} = \mathcal{GC}(\theta_n, s_{nk}) \qquad \forall k \tag{1}$$

The next step is to find possible trips from these stop locations which go in the direction of the next tag location. For each stop s_{nk}, find the possible trips $\mathcal{T}_{nk} = \{tr_{kl}, l = 1, 2, ...\}$

which are within τ minutes of tag time t_n assuming that bus can be late or early on a given stop s_{nk} by τ minutes. This delay parameter τ is flexible and can be adjusted for the given algorithm. With greater value of τ, more trip options will be created. This will obviate the problem of incorrect sub-route (Section 3.2) trip being selected. Then we calculate the delay for different trips as:

$$\Delta_{kl} = |t_{tr_{kl}} - t_n| \qquad \forall k, l \tag{2}$$

Using the trip information, for each trip l, find a set of alighting stops $A_{nkl} = \{a_{klm}, m = 1, 2, ...\}$ which is within ϵ miles of next tag location θ_{n+1}. Again, ϵ is flexible and can be assumed as any suitable value. This will avoid the problem of finding wrong alighting stop mentioned in Munizaga and Palma 2012. Let $\mathcal{IV}_{klm}$ be the in-vehicle time for the trip tr_{kl} with alighting stop a_{klm} and w_{klm} be the walking distance from alighting location a_{klm} to the next tag location θ_{n+1}. All the potential stops and trips can be connected via a graph shown in Figure 3.

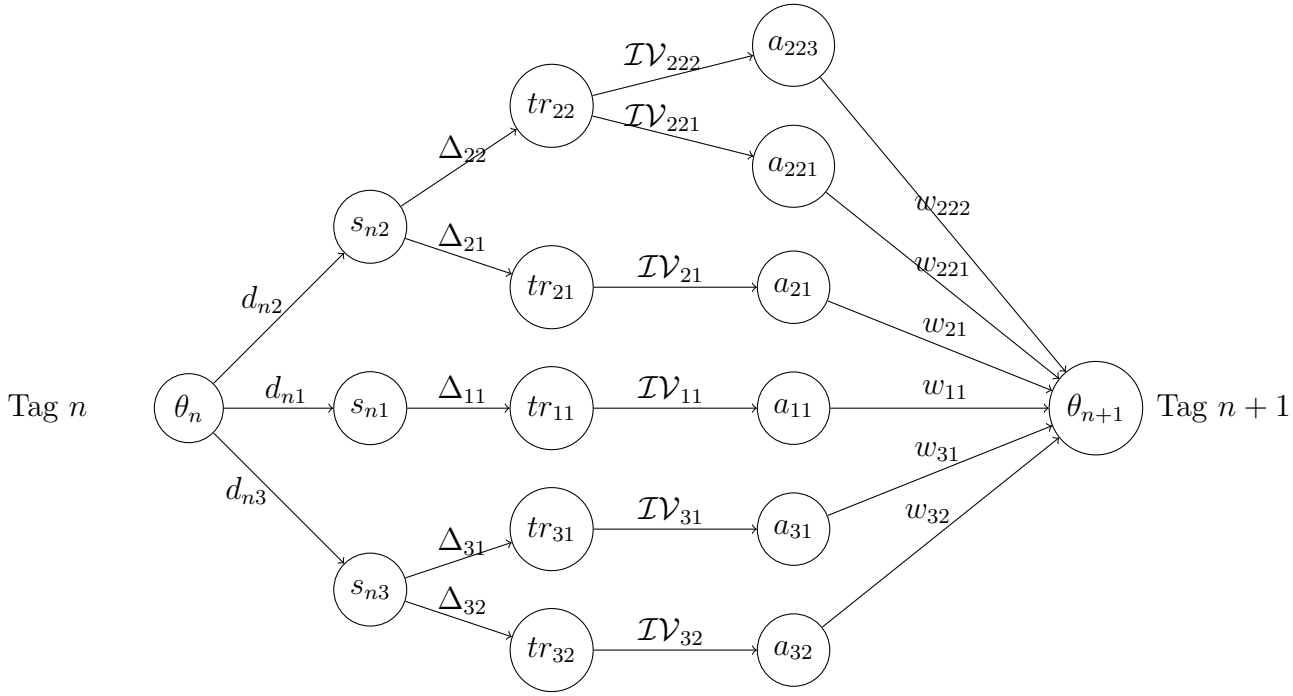


Figure 3: Network of possible trips

4.2. Probability calculation for possible trips

Let $P(s_{nk})$ be the probability of boarding stop s_{nk} from tag location θ_n. This probability is a function of great circle distance d_{nk} which is created because of the GPS inaccuracy and can be modeled as a zero mean Gaussian distribution (van Diggelen 2007), given as:

$$P(s_{nk}) = f(\sigma_k, d_{nk}) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp^{-0.5(\frac{d_{nk}}{\sigma_k})^2} \qquad \forall k \tag{3}$$

If we assume s_{nk} was the actual boarding location, then d_{nk} is an estimate of the magnitude of GPS error. The standard deviation of these values, i.e. σ_k, is our estimate of the

GPS error. We estimate σ_k using the median absolute deviation, which is a robust estimator of standard deviation. The value of σ_k can be given as:

$$\sigma_k = 1.4826 * \text{median}(d_{nk}) \qquad \forall k \tag{4}$$

The probability of taking a trip tr_{kl} from stop s_{nk}, i.e. $P(tr_{kl}|s_{nk})$, is a function of bus delay Δ_{kl}:

$$P(tr_{kl}|s_{nk}) = f(\Delta_{kl}) \qquad \forall k, l \tag{5}$$

The probability distribution function $f(\Delta_{kl})$ of bus delay can be calculated using APC-VL data, which contains vehicle arrival times on limited stops for a given bus route trip l. We can model the probability of reaching the next tag location θ_{n+1} by taking trip tr_{kl} and alighting at stop a_{klm} using a multinomial logit route choice model given as:

$$P(a_{klm}|tr_{kl}, s_{nk}) = \frac{\exp^{-(\beta_1 \mathcal{IV}_{klm} + \beta_2 \frac{w_{klm}}{s})}}{\sum_{p,g} \exp^{-(\beta_1 \mathcal{IV}_{kpg} + \beta_2 \frac{w_{kpg}}{s})}} \quad \forall l, k \tag{6}$$

where, s is the walking speed which is assumed as 3.0 miles per hour. β_1 and β_2 are the parameters which shows the disutility of walking in comparison to in-vehicle travel time according to user behavior.

Finally, assuming the random variables describing the probability distributions are independent, we can evaluate the probability of traversing from location θ_n to θ_{n+1} using any of the trips by multiplying (3), (5) and (6) which is the product of the following components.

- GPS inaccuracy of the current tag
- Bus delay of the current tag
- Route choice model consisting of in-vehicle and walking time between the current tag and the next tag.

$$\begin{aligned} P(a_{klm}, tr_{kl}, s_{nk}|\theta_n, \theta_{n+1}) &= P(a_{klm}|tr_{kl}, s_{nk}, \theta_n, \theta_{n+1}) P(tr_{kl}|s_{nk}, \theta_n, \theta_{n+1}) P(s_{nk}|\theta_n, \theta_{n+1}) \\ &= f(\sigma_k, d_{nk}) f(\Delta_{kl}) P(a_{klm}|tr_{kl}, s_{nk}) \quad \forall l, k, m \end{aligned} \tag{7}$$

Hence, the most likely boarding and alighting stops for this tag n can be inferred using the trip for which $P(a_{klm}, tr_{kl}, s_{nk}|\theta_n, \theta_{n+1})$ is maximum.

4.3. Extension to pay-exit cases

If there is a combination of pay-exit and regular tags (Section 3.3), then the probability calculations change according to available information. These cases are discussed below:

4.3.1. Current tag is pay exit and next tag is regular

In this case, the probability of each trip consists of three components:

- GPS inaccuracy of the current tag
- Bus delay of the current tag
- Route choice model consisting of only walking time between the current tag and the next tag.

The final expression is given below:

$$P(a_{klm}, tr_{kl}, s_{nk}|\theta_n, \theta_{n+1}) = f(\sigma_k, d_{nk}) f(\Delta_{kl}) \frac{\exp^{-(\beta_2 \frac{w_{klm}}{s})}}{\sum_{p,g} \exp^{-(\beta_2 \frac{w_{kpg}}{s})}} \quad \forall l, k \tag{8}$$

4.3.2. Current tag is regular and next tag is pay exit

For this case, if two different routes are used for making these two trips, then the probability of each alternative to go from the current boarding to the next alighting consists of three components:

- GPS inaccuracy of the current tag and the next tag
- Bus delay of the current tag and the next tag
- A common route choice model consisting of in-vehicle travel time of the two trips and the walking time between the trips.

The final expression is given below:

$$P(a_{klm^{12}}, tr_{kl^1}, tr_{kl^2}, s_{nk^1}, s_{nk^2}|\theta_n, \theta_{n+1}) = f(\sigma_k^1, d_{nk}^1) f(\sigma_k^2, d_{nk}^2) f^1(\Delta_{kl}^1) f^2(\Delta_{kl}^2) \frac{\exp^{-(\beta_1 \mathcal{IV}_{klm^1} + \beta_1 \mathcal{IV}_{klm^2} + \beta_2 \frac{w_{klm^{12}}}{s})}}{\sum_{g,p^1,p^2} \exp^{-(\beta_1 \mathcal{IV}_{kpg^1} + \beta_1 \mathcal{IV}_{kpg^2} + \beta_2 \frac{w_{kpg^{12}}}{s})}} \quad \forall l, k \tag{9}$$

If both tags use the same or parallel routes, we can make use of APC data to assign the alighting of the current tag and boarding of the next tag. Usually some particular stops at the end of the routes are more common stops for alighting. Using route information, we calculate the proportion of alighting at these stops for each route, then assign the required boarding and alighting stops proportionally for each case in the AFC data. In this way, we may not get exact inference in the individual level, but on an aggregate level, the results will be consistent. Anyhow, the percentage of these cases in the AFC database is very low.

4.3.3. Current tag is pay exit and next tag is pay exit

In this case, the probability of each trip consists of three components

- GPS inaccuracy of the next tag
- Bus delay of the next tag

- route choice model consisting of in-vehicle travel time and walking time of the next trip.

The final expression is given below:

$$P(a_{klm}, tr_{kl}, s_{nk}|\theta_n, \theta_{n+1}) = f(\sigma_k, d_{n+1,k}) f(\Delta_{kl}) f(\frac{\exp^{-(\beta_1 \mathcal{IV}_{klm} + \beta_2 \frac{w_{klm}}{s})}}{\sum_{p,g} \exp^{-(\beta_1 \mathcal{IV}_{kpg} + \beta_2 \frac{w_{kpg}}{s})}}) \quad \forall l, k \qquad (10)$$

4.4. Transfer detection

Transfer information given in the AFC data may not be reliable. Consistent with the fair policy, the AFC system considers a tag as a transfer if it has been made within 150 minutes of the previous tag time. The method described in Nassir et al. 2011 is used to detect transfers. The method infers next tag as transfer if it has been made within 30 minutes and boarding if it has been made after 90 minutes of alighting. Between 30 and 90 minutes, after alighting at a station, the walking time (W) and setback delay time (D) (due to possible minor activities like buying coffee or newspaper) is considered and a time t_{acc} is calculated which is the time when boarding stop becomes accessible. Then, the number of opportunities (N_{opp}) to catch the next bus is calculated between the time t_{acc} and the actual boarding time of the next tag by counting the number of trips in GTFS data within the time range. If $N_{opp} \leq 1$, we infer the next tag as transfer, otherwise, there is a possibility of an activity and we mark the next tag as boarding.

Complete trip chaining algorithm is described in Algorithm 1.

Algorithm 1 Robust Trip Chaining Algorithm

procedure
  data structure
  n: an AFC tag
  pe: 1, if tag is pay exit, 0, otherwise
  seq: sequence number of the tag serial number for the given date
  ser: sequence number of a transit stop for a given tripID in GTFS data
  P: list of possible stops around tag location
  L: list of possible trips for a given stop
  All other notations are consistent with *Table 1*
  function FindPossibleStops($tag[n]$)
    $P \leftarrow []$
    $st\_list \leftarrow$ find a list of stops for $tag[n].r$ and $tag[n].\delta$ from GTFS
    for each stop s in st_list **do**
      if $dist(s, tag[n].\theta) < \alpha$ **then**
        append s to P
    return P
  function FindPossibleTrips(p)
    $L \leftarrow []$
    $tr\_list \leftarrow$ find all the trips for given stop $p.r$, $p.\delta$ from GTFS
    for each trip l in tr_list **do**
      if $abs(l.dep - tag[n].t) \leq \tau$ **then**
        append l to L
    return L
  function InferBoardingAlighting(l, $tag[n]$, $tag[n+1]$)
    if the inference is for alighting **then**
      $al\_stops \leftarrow$ find stops with stop sequence greater than $l.ser$
      return alighting stops within distance ϵ of the $tag[n+1]$
    else
      $bo\_stops \leftarrow$ find stops with stop sequence less than $l.ser$
      return boarding stops within distance ϵ of the $tag[n]$
Algorithm
  for each n **do**
    $Prob \leftarrow []$
    if $tag[n].seq =$ last tag of the day **then**
      take $tag[n+1] =$ first tag of the day for that serial number
    $P \leftarrow$ FindPossibleStops($tag[n]$)
    for each stop p in P **do**
      $L \leftarrow$ FindPossibleTrips(p)
      for each trip l in L **do**
        Depending on $tag[n].pe$ and $tag[n+1].pe$
        $L \leftarrow$ InferBoardingAlighting(l, $tag[n]$, $tag[n+1]$)
        Calculate $Prob[l]$
    Find the trip with maximum probability
    Infer the boarding and alighting of $tag[n]$ and $tag[n+1]$ based on that trip

5. Data Description and Preparation

5.1. Automated Data

Metro Transit is the primary transit agency in the Twin Cities, offering an integrated network of buses, light rail and commuter trains. The automated data used in this study is collected by Metro Transit. GTFS, AFC and APC-VL data are required for this research. These datasets were uploaded to the PostgreSQL server and queried using R package RPostgreSQL (Conway et al. 2017). A brief discussion of different types of data and their preparation is given below:

5.1.1. Automatic Fare Collection (AFC) Data

The AFC data used for this research comes from the University of Minnesota student transit pass (U-Pass) data. The AFC system records the fare related information when a passenger pays for a trip. This includes a particular serial ID assigned to the pass, date and time of the tag, route information, geographical coordinates of the tag, transfer information, etc. A sequence column was added to the data which keeps track of the sequence of the tags made by a passenger on a particular day. Pay-exit column was also added to the data by checking the buses and their direction in which they are pay-exit. Several issues with data were resolved before running the trip chaining algorithm. For example, AFC data for light rail does not have geographical coordinates but contains the station information where the passenger boarded the light rail, in which case we do not have to search for possible boarding stops. Another issue is that light rail AFC data does not have direction information. This is because light rail stations serve the trains in both directions. We inferred the direction of light rail trips using the next tag location.

After the initial data processing, there are still some tags which do not have any geographic information. These mainly consist of the buses not operated by Metro Transit (e.g operated by Minnesota Valley Transit Authority (MVTA), First Transit, etc). We removed such entries for the analysis because the GTFS data was unavailable for these services. The data also contains some tags which have geographic location outside the transit service region, so we removed such entries from the dataset. We also removed the cases where a single tag is made by a passenger on a day as trip chaining requires at least two trips made by a passenger in order to estimate the origin and destination. Table 2 shows the number of tags in the data set for four typical weekdays (March 07, 2016 to March 10, 2016).

Table 2: Tag Description

Description	**Number of tags**	**Percentage**
Total tags	85,456	
Missing geographical coordinates	4,785	5.6
Outlier geographical location	3,515	4.1
Single tags	10,782	12.6
Total remaining tags	66,374	77.7

5.1.2. General Transit Feed Specification (GTFS) Data

GTFS (Google 2005) data contains schedule information of the buses and light rail, including their stops location, route information, scheduled arrival and departure time, etc. For trip chaining, we selected the appropriate service ID for the study period and then query the data.

5.1.3. Automatic Passenger Count-Vehicle Location (APC-VL) Data

The automatic passenger count system records date, time, transit route, stop and trip information, departure and arrival time at time point stops, number of boarding and alighting at every stop, and geographical coordinates of stops.

5.2. Model calibration

The probability distribution functions required for the trip chaining algorithm were prepared as follows:

5.2.1. Gaussian model for GPS inaccuracy

To calibrate (equation (3-4)), we created a list of the AFC tag locations for which only one stop is found within a buffer distance of 0.1 miles and calculated the values of the d_{nk}. These stops can be regarded as ground truth data required for calibration. Using these values, we calculated the value of $\sigma_k = 55.25$ feets.

5.2.2. Bus delay probability distribution

As mentioned before, automatic APC-VL data contains bus arrival time at limited stops. We used the available arrival times to calculate the probability of bus route being early or late. We used a discrete distribution for the bus delay distribution (equation (5)) with a class range of one-minute intervals.

5.2.3. Route choice model

For (equation (6)), we assumed the value of $\beta_1 = 1$, $\beta_2 = 2$, and the walking speed, $s = 3$ miles per hour for our route choice model. These values are consistent with the literature (Hunt 1990, Guo and Wilson 2007, Raveau et al. 2012).

6. Results

6.1. Analysis of the results

After data preparation, Algorithm 1 was implemented in R (R Core Team 2017) for U-Pass (University of Minnesota Pass) AFC data from March 07, 2016 to March 10, 2016. Figure 4 shows the number of trips made by the U-Pass holders during the analysis period. We can observe the morning peak between 6:30 A.M. to 9:30 P.M. and afternoon peak between 3:00 P.M. to 6:30 P.M.

After removing all the outliers described above, 66,374 out of 85,456 tags were left. Out of remaining 66,374 tags, both origin and destination of 56,423 (85%) tags were successfully inferred in comparison to 46,507 (70%) tags being inferred using the baseline algorithm described in Nassir et al. 2011. Table 3 summarizes the results in which about 81% of pay exit cases were inferred using the proposed algorithm in comparison to no inference using

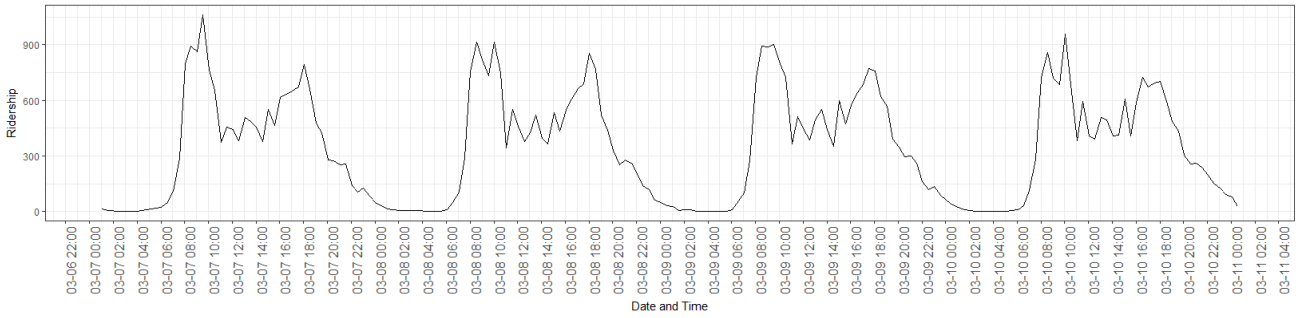


Figure 4: Time distribution of the trips in U-Pass data

the baseline algorithm. Another comparison was done between the two algorithms for inferred boarding and alighting. Out of 46,507 inferred regular cases, 384 (0.8%) boardings and 300 (0.6%) alightings were different. About 9% of the tags were inferred as transfers in comparison to 17% in the original AFC data which considers every tag as a transfer if it is made within 2 hours and 30 minutes of the previous tag. One point of interest is whether the last tag of the day can be inferred using the first tag of the day. We found that out of 26,275 last tags, the algorithm is able to infer the boarding and alighting of 21,110 tags (80%). This shows that this assumption works well in practice. Among the tags which are not inferred, about 59% are not inferred because no stop was found within walking distance from the current alighting location to the next boarding location. The likely reason for this non-inference is the use of another mode of transportation between two transit trips. We also observed that due to wrong selection of trip IDs from GTFS data, around 558 tags were not inferred using the baseline algorithm because the boarding time of the next tag was less than the alighting time of the current tag. The proposed algorithm eliminated this problem. This is because of the consideration of a list of possible trajectories for a given tag in the proposed algorithm in comparison to only one trip in the baseline algorithm.

Table 3: Comparison of the results between the baseline and the proposed method

Algorithm	Baseline Method	Proposed Method	Percent Improvement
Pay Exit Count	5,562	5,562	
Regular Count	60,812	60,812	
Pay Exit Inferred	0	4,504	7%
Regular Inferred	46,507	51,919	8%
Total Tag Count	66,374	66,374	
Total tags inferred	46,507 (70%)	56,423 (85%)	15%

Note: The percentage improvement is calculated with respect to the total number of tags (i.e. 66, 374)

The selection of the most likely trajectory based on the highest probability may result in accumulation of the inference error if there are multiple likely trajectories instead of a dominant one. In order to check for this possibility, we calculated the percentage difference between the probabilities of the first and the second (if exists) most likely trajectories for

every tag. The percentage difference is calculated with respect to the highest probability. A histogram of the percentage difference of these probabilities is shown in Figure 5. We found that more than 95% of the values were greater than 19% difference. To test if there exist a significant number of trips with multiple likely trajectories, we extracted 5% of the trips from lower tail of the distribution (shown by the dashed line) to compare the means of the probabilities of the first and the second most likely trajectories. We used the paired two sample T-test to compare the means.

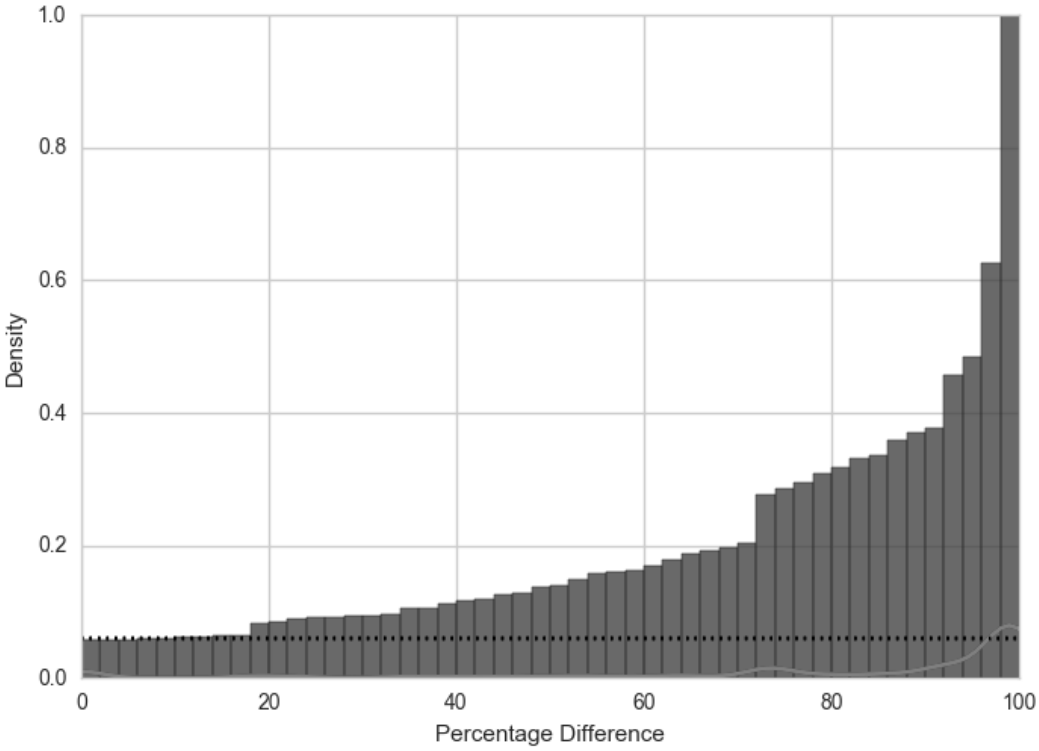


Figure 5: Distribution of the percentage difference between the probabilities of the first and the second (if exists) most likely trajectories

$$\begin{aligned} H_0 &: \mu_{\text{first}} = \mu_{\text{second}} \\ H_1 &: \mu_{\text{first}} \neq \mu_{\text{second}} \end{aligned} \tag{11}$$

We found a T-statistic value of 24.383 which is greater than the critical value at 99% confidence level. This rejects the null hypothesis that the means of the probabilities of the first and second most likely trajectories are equal. We recommend to perform this test to check the quality of the results. If there exists a significant number of trips with multiple likely trajectories, then we either should consider all the likely trajectories for that tag or choose a trajectory randomly from the set of likely trajectories.

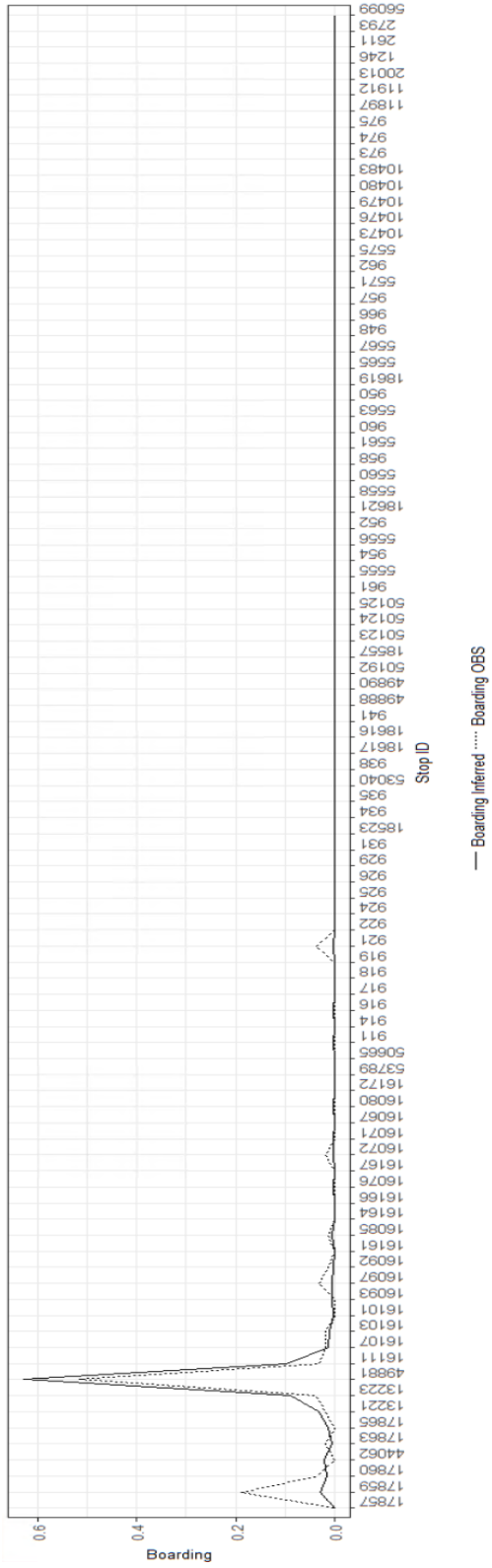
It is difficult to validate the trip chaining results for an open transit system because of the lack of ground truth data available to compare the results. We use the transit on-board

survey data from 2016 to compare the total number of boarding and alighting on different stops of a route. The on-board survey (OBS) collects data from individuals about their travel itinerary such as origin and destination of the trip, boarding and alighting stops, route, transfer information, etc. Then an expansion factor is used to expand the survey for the total boarding and alighting counts obtained from the APC data. We analyzed the high ridership routes such as route 2, 3 and Metro Green Line for this purpose. The overall proportion of the boarding and alighting on different stops of these routes were similar. The results for route 3 in eastbound direction is presented in Figure 6. We can observe that the boarding proportions (Figure 6(a)) are almost similar at every stop except few stops. Figure 6(b) shows the comparison of alighting proportion at different stops. The pattern in alighting looks similar but the difference is quite high for some of the stops. We believe that the error in the boarding and alighting proportions is caused by the low sampling rate and possibly inaccurate boarding and alighting stops from the on-board survey. Wang et al. 2011 also faced similar challenges to use OBS for validation purposes. We also compared the number of transfers made by the passengers to assess the accuracy of transfer inference. We found the proportion of transfers similar to on-board survey. For example, for route 3 eastbound, the results shows 3.6 % transfers using the proposed algorithm in comparison to 3.5 % and 10.3 % using the on-board survey data and the AFC system respectively.

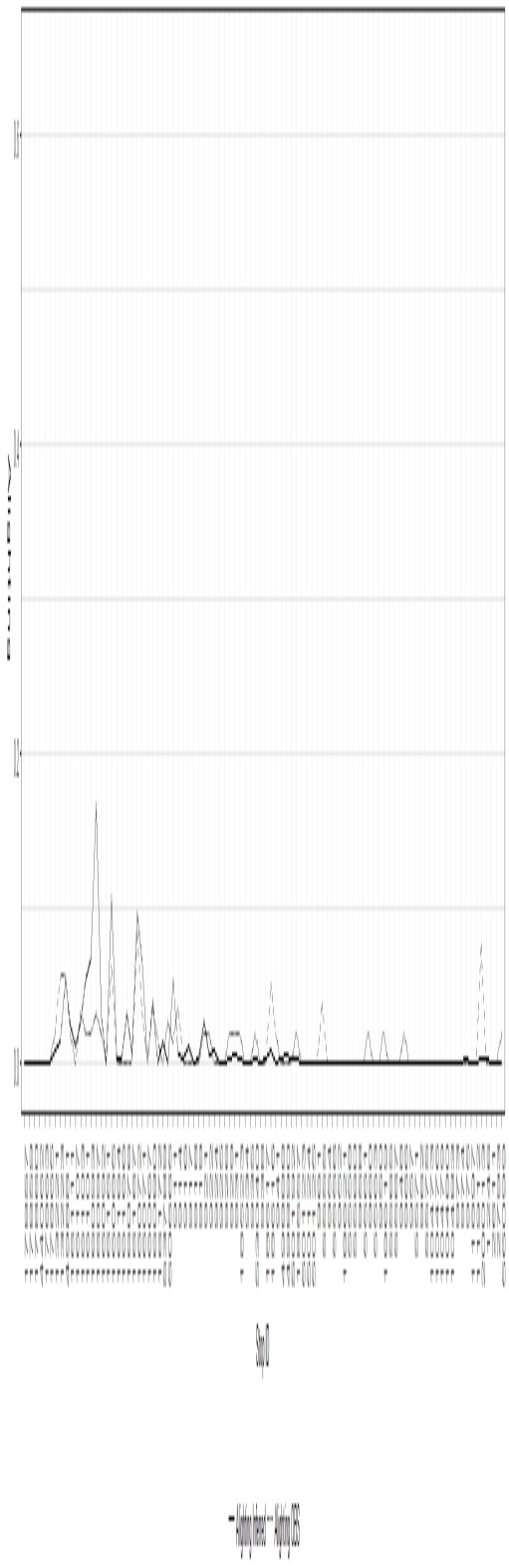
6.2. Applications using the inferred results

To summarize the outputs, heat maps of trip origins and destinations are prepared (Figure 7). The maps show that during morning peak hours, most of the trips originate from the areas east of the campus, Downtown and southwest Minneapolis, Downtown St. Paul, area around the university campus and Metro Green Line, while trip destinations are mainly at the university campus. Looking at the results for the evening peak hours, the origins and destinations look reversed, where most trips begin from the university campus and end at popular morning origin locations.

We compared the route ridership to assess the most common transit routes used by university students. Table 4 shows the high ridership routes and stops. In this table, as expected Metro Green Line has the highest ridership as it connects Downtown Minneapolis and Downtown St. Paul via university campus through two stations, East Bank Station and West Bank Station, which are also the popular locations for boarding and alighting in the stop table. Route 2 and route 3 are the most common bus routes used by the university students who live close to the campus. Route 3 connects Downtown Minneapolis and Downtown St. Paul via university by serving areas around the campus. Route 6, route 114 and route 113 serve the southwest suburbs while route 465 and 87 serve the southern suburbs. It is interesting to see that many students from suburbs use bus to commute to the campus. In the stop table (Table 4), stops located in the university campus such as East Bank Station, Pleasant Street & Jones Hall, West Bank Station, Washington Avenue & Coffman Union and Washington Avenue & Oak Street SE show high ridership. Other high ridership stops shown in the table are Metro Green Line stations. Finally, 15th Avenue SE and Como Avenue is also a popular stop for boarding and alighting served by route 3.

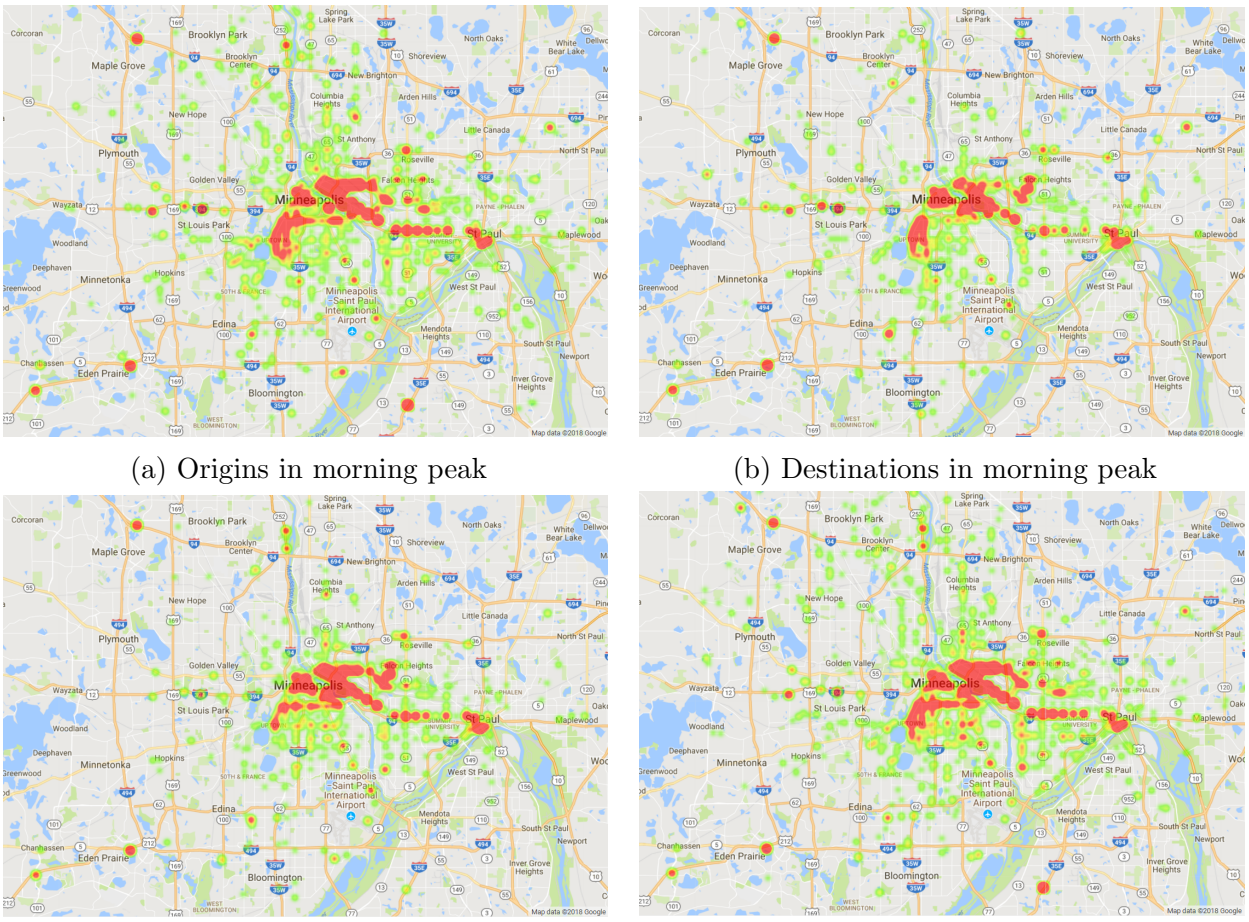


(a) Comparison of route 3 eastbound boarding proportions



(b) Comparison of route 3 eastbound alighting proportions

Figure 6: Comparison of boarding and alighting proportions from on-board survey and inferred results for route 3 in eastbound direction



(a) Origins in morning peak

(b) Destinations in morning peak

(c) Origins in evening peak

(d) Destinations in evening peak

Figure 7: Intensity of trip origins and destinations. (For interpretation of colors in this figure, the reader is referred to the web version of this article.)

The highest number of tags was made on the Metro Green Line stations for which we did stop level origin-destination analysis. In Figure 8(a), we can observe that in the morning peak and eastbound direction, most trips start from Downtown Minneapolis at the western end of the line to the East Bank and West Bank Stations on the university campus or from Downtown St. Paul Union Depot (Figure 8(b)) at the eastern end of the line to the East Bank Station. Most of the students commute from the stations east of campus, for example Stadium Village, Prospect park and Westgate which are closer to the university. Conversely, during the evening peak, most trips go from East Bank and West Bank Stations to the popular origin locations in the morning (Figure 8(c) and Figure 8(d)).

6.3. Discussion

In this section, we discuss the possible ways to infer the non-inferred tags. The proposed method infers the boarding and alighting of the tags made by the passenger during the day based on the assumptions given in section 2. If these assumptions are not satisfied, then it cannot infer the boarding and alighting location of a given tag. Such trips (tags) are called unlinked trips (He and Trépanier 2015). The inference of such trips is possible using a

Table 4: Routes and stop locations with high ridership

Route	Ridership
Metro Green Line	22,144
3	12,213
2	6,340
6	3,014
465	2,274
114	1,569
113	1,207
901	1,126
87	1,073
698	794

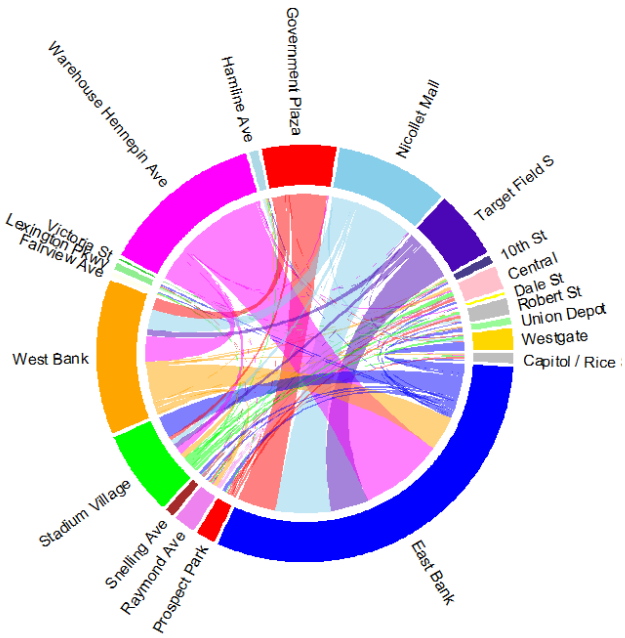
Stop/Station	Boarding	Alighting
East Bank Station & Platform	7,052	7,314
Pleasant St & Jones Hall	3,423	3,265
Stadium Village Station & Platform	2,928	2,783
West Bank Station & Platform	2,924	2,723
Washington Ave & Coffman Union	1,637	2,006
Westgate Station & Platform	1,441	1,280
15th Ave SE & Como Ave SE	1,105	1,013
Washington Av & Oak St SE	971	971
Prospect Park Station & Platform	923	828
Warehouse Hennepin Ave Station & Platform	714	626

method proposed by He and Trépanier 2015, which assumes that passengers tend to follow the same routine, and the historical alighting location and time information can be used to infer the alighting location of an unlinked trip. The method extracts the historical destinations for a passenger and tries to estimate the probability of alighting on these locations. The probability is found using spatial and temporal proximity of the historical alighting and the potential alighting. The method can be used in our case for the regular tags. We need to repeat the procedure of finding the spatial and the temporal probabilities for all the possible trajectories found for a given tag. However, the method may not be useful for the pay-exit cases. For example, for a commuter who takes a regular route in the morning and pay-exit route in the evening, there will be no historical alighting and boarding location for the current and the next tag location respectively. Another disadvantage of combining the method proposed by He and Trépanier 2015 and the proposed method is heavy computational time as the spatial and temporal probabilities need to be calculated for each possible trajectory.

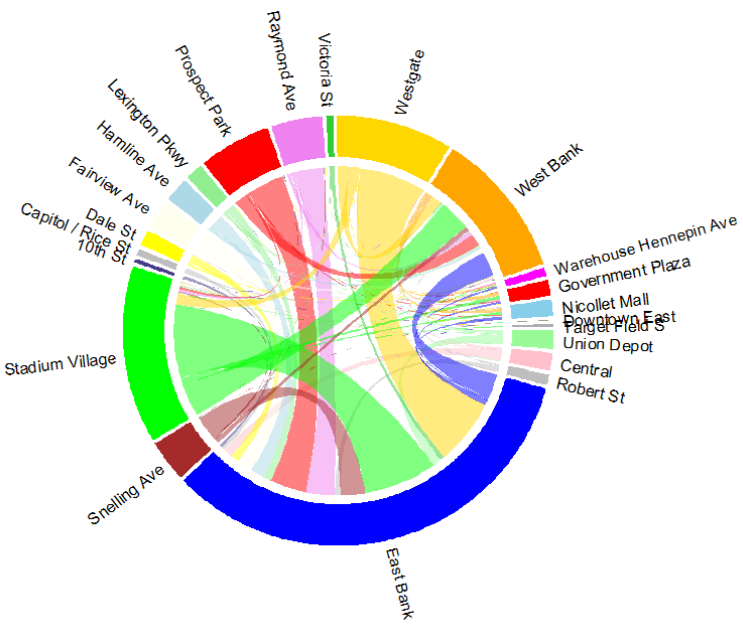
Transit agencies require full O-D matrix for all the trips made by users given the errors and the missing information. This can be achieved using the boarding and alighting count data available from APC data. The O-D matrix obtained from AFC data using trip chaining algorithm can be used as a seed or prior matrix in optimization methods proposed by Van Zuylen and Willumsen 1980 or Spiess 1987. These optimization methods promise to perform better with a good quality seed matrix, which we can obtain from the trip chaining results. Another possibility is to proportionally assign the non-inferred boarding and alighting based on the APC data. Although these methods may not infer the correct boarding and alighting on an individual level, they will improve the results on an aggregate level.

7. Conclusions and Recommendations for Future Research

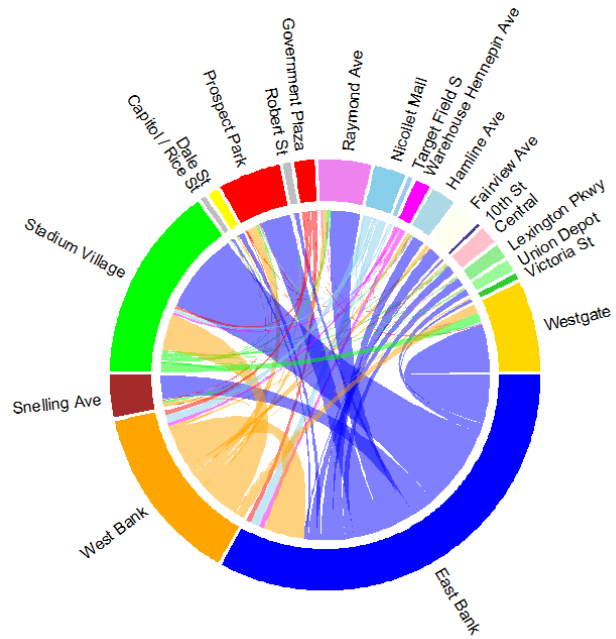
This research proposes a robust method for trip chaining of transit smart card data, which tries to relax various assumptions on the parameters used in the existing trip chaining algorithms. The parameters can vary according to the quality of data and user behavior in different transit systems, so a fixed value cannot be assumed for different transit systems.



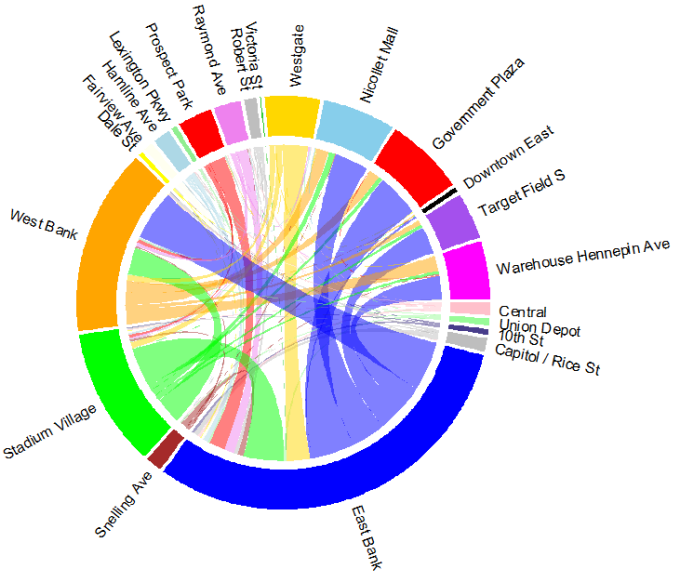
(a) Flow of passengers in the morning peak in the eastbound direction



(b) Flow of passengers in the morning peak in the westbound direction



(c) Flow of passengers in the evening peak in the eastbound direction



(d) Flow of passengers in the evening peak in the westbound direction

Figure 8: Passenger origin-destination flow on Metro Green Line light rail. (For interpretation of colors in this figure, the reader is referred to the web version of this article.)

This is evident from trip chaining results for the Twin Cities AFC data. The proposed method provides the flexibility to assume a higher value for these parameters to avoid wrong

inference of origin and destination.

The method uses probability distributions for potential boarding stop location, bus delay and passenger's route choice behavior. By combining these probabilities, it infers the most likely trajectory of the passenger. Though being an open transit system with pay-exit buses and sub-routes, these attributes create various problems for trip chaining. Using the proposed method, various problems such as erroneous GPS locations, selection of wrong trip for inference, and pay-exit cases are addressed. The proposed algorithm can also be suitably modified to deal with different pay exit cases.

The O-D matrix results can be used in multiple ways to understand the travel behavior of passengers in a transit system. We presented the ridership analysis on an aggregate level for the Twin Cities and also the route level analysis for a light rail transit line. We can also use the trip chaining results by creating clusters of customers based on their regularity in using transit system. These results can inform planners for better decisions to improve transit services.

Current research can be expanded in multiple directions. The case where the current tag is regular and the next tag is pay-exit and both tags use the same route is analyzed using a method of proportions. Additional information from other data sources can help in development of a suitable algorithm for this case. The results obtained from trip chaining can be used for other research such as trip purpose inference, analyzing spatial and temporal travel pattern, route choice behavior analysis of passengers and transit assignment models.

Acknowledgement

This research is conducted at the University of Minnesota Transit Lab, currently supported by the following, but not limited to, projects:

- National Science Foundation, award CMMI-1637548
- Minnesota Department of Transportation, Contract No. 1003325 Work Order No. 15
- Minnesota Department of Transportation, Contract No. 1003325 Work Order No. 44
- Transitways Research Impact Program (TIRP), Contract No. A100460 Work Order No. UM2917

The authors are grateful to Metro Transit for sharing the data. We are also grateful to the anonymous referees for their constructive input to improve the quality of this article. Any limitation of this study remains the responsibility of the authors.

Alfred Chu, K. and Chapleau, R. 2008, 'Enriching Archived Smart Card Transaction Data for Transit Demand Modeling', *Transportation Research Record: Journal of the Transportation Research Board* (2063), 63–72.

Alsger, A., Assemi, B., Mesbah, M. and Ferreira, L. 2016, 'Validating and improving public transport origin-destination estimation algorithm using smart card fare data', *Transportation Research Part C: Emerging Technologies* **68**, 490–506.

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L. and Hickman, M. 2018, 'Public transport trip purpose inference using smart card fare data', *Transportation Research Part C: Emerging Technologies* **87**(January), 123–137.

Attanucci, J. & Wilson, N. 1981, 'Bus Transit Monitoring Manual: Volume 1: Data Collection Program Design', *US Department of Transportation* **1**.

Barry, J., Freimer, R. and Slavin, H. 2009, 'Use of Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City', *Transportation Research Record: Journal of the Transportation Research Board* **2112**, 53–61.

Barry, J. J., Newhouser, R., Rahbee, A. and Sayeda, S. 2007, 'Origin and Destination Estimation in New York City with Automated Fare System Data', *Transportation Research Record: Journal of the Transportation Research Board* **1817**(1), 183–187.

Briand, A. S., Côme, E., Trépanier, M. and Oukhellou, L. 2017, 'Analyzing year-to-year changes in public transport passenger behaviour using smart card data', *Transportation Research Part C: Emerging Technologies* **79**, 274–289.

Chu, K. and Chapleau, R. 2010, 'Augmenting Transit Trip Characterization and Travel Behavior Comprehension', *Transportation Research Record: Journal of the Transportation Research Board* **2183**, 29–40.

Conway, J., Eddelbuettel, D., Nishiyama, T., Prayaga, S. K. and Tiffin, N. 2017, 'RPostgreSQL: R Interface to the 'PostgreSQL' Database System'.
URL: *https://cran.r-project.org/package=RPostgreSQL*

Farzin, J. 2008, 'Constructing an Automated Bus Origin-Destination Matrix Using Farecard and Global Positioning System Data in São Paulo, Brazil', *Transportation Research Record: Journal of the Transportation Research Board* **2072**(2072), 30–37.

Google 2005, 'General Transit Feed Specification'.
URL: *http://code.google.com/ transit/spec/transit_feed_specification.htm*

Gordon, J. B., Koutsopoulos, H. N. and Wilson, N. H. 2018, 'Estimation of population origin–interchange–destination flows on multimodal transit networks', *Transportation Research Part C: Emerging Technologies* **90**, 350 – 365.

Gordon, J., Koutsopoulos, H., Wilson, N. and Attanucci, J. 2013, 'Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data', *Transportation Research Record: Journal of the Transportation Research Board* **2343**, 17–24.

Guo, Z. and Wilson, N. H. 2007, 'Modeling effects of transit system transfers on travel behavior: case of commuter rail and subway in downtown boston, massachusetts', *Transportation Research Record* **2006**(1), 11–20.

He, L. and Trépanier, M. 2015, 'Estimating the Destination of Unlinked Trips in Transit Smart Card Fare Data', *Transportation Research Record: Journal of the Transportation Research Board* **2535**, 97–104.

Hunt, J. D. 1990, 'A Logit Model of Public Transport Route Choice', *ITE Journal* **December**, 26–30.

Ingvardson, J. B., Nielsen, O. A., Raveau, S. and Nielsen, B. F. 2018, 'Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: A smart card data analysis', *Transportation Research Part C: Emerging Technologies* **90**(September 2017), 292–306.

Khani, A. 2018, 'Transit Demand Analysis and User Classification Using Automatic Fare Collection (AFC) Data', *TREC Friday Seminar Series* **144**.
URL: *https://pdxscholar.library.pdx.edu/trec_seminar/144*

Kim, J., Corcoran, J. and Papamanolis, M. 2017, 'Route choice stickiness of public transport passengers: Measuring habitual bus ridership behaviour using smart card data', *Transportation Research Part C: Emerging Technologies* **83**, 146–164.

Kumar, P., Khani, A. and He, Q. 2018, A Probabilistic Trip Chaining Algorithm for Transit Origin-Destination Matrix Estimation Using Automated Data, *in* 'Transportation Research Board Annual Meeting 2018'.

Kusakabe, T. and Asakura, Y. 2014, 'Behavioural data mining of transit smart card data: A data fusion approach', *Transportation Research Part C: Emerging Technologies* **46**, 179–191.

Lee, S. G. and Hickman, M. 2014, 'Trip purpose inference using automated fare collection data', *Public Transport* **6**(1-2), 1–20.

Li, J. Q. 2012, 'Match bus stops to a digital road network by the shortest path model', *Transportation Research Part C: Emerging Technologies* **22**, 119–131.

Li, T., Sun, D., Jing, P. and Yang, K. 2018, 'Smart Card Data Mining of Public Transport Destination: A Literature Review', *Information* **9**(1), 18.

Luo, D., Cats, O. and van Lint, H. 2017, 'Constructing Transit Origin–Destination Matrices with Spatial Clustering', *Transportation Research Record: Journal of the Transportation Research Board* **2652**, 39–49.

Ma, X.-l., Wang, Y.-h., Chen, F. and Liu, J.-f. 2012, 'Transit smart card data mining for passenger origin information extraction', *Journal of Zhejiang University Science C* **13**(10), 750–760.

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

Munizaga, M. A. and Palma, C. 2012, 'Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile', *Transportation Research Part C: Emerging Technologies* **24**, 9–18.

Nassir, N., Khani, A., Lee, S., Noh, H. and Hickman, M. 2011, 'Transit Stop-Level Origin-Destination Estimation Through Use of Transit Schedule and Automated Data Collection System', *Transportation Research Record: Journal of the Transportation Research Board* **2263**, 140–150.

Navick, D. and Furth, P. 2002, 'Estimating passenger miles, origin-destination patterns, and loads with location-stamped farebox data', *Transportation Research Record: Journal of Journal of the Transportation Research Board* **107-113**(02), 2466.

Navy, R. 2008, 'Admiralty manual of navigation: The principles of navigation, volume 1'.

Newson, P. and Krumm, J. 2009, 'Hidden Markov Map Matching Through Noise and Sparseness', *17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* pp. 336–343.

Pelletier, M. P., Trépanier, M. and Morency, C. 2011, 'Smart card data use in public transit: A literature review', *Transportation Research Part C: Emerging Technologies* **19**(4), 557–568.

Perrine, K., Khani, A. and Ruiz-Juri, N. 2015, 'Map-Matching Algorithm for Applications in Multimodal Transportation Network Modeling', *Transportation Research Record: Journal of the Transportation Research Board* **2537**(2537), 62–70.

R Core Team 2017, 'R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria'.
URL: *https://www.r-project.org/*

Raveau, S., Guo, Z., Muñoz, J. C. and Wilson, N. H. M. 2012, 'Route Choice Modelling on Metro Networks', *Conference on Advanced Systems for Public Transport* (56 2), 1–13.

Robinson, S., Narayanan, B., Toh, N. and Pereira, F. 2014, 'Methods for pre-processing smartcard data to improve data quality', *Transportation Research Part C: Emerging Technologies* **49**, 43–58.

Seaborn, C., Attanucci, J. and Wilson, N. H. M. 2009, 'Using Smart Card Fare Payment Data To Analyze Multi- Modal Public Transport Journeys in London', *Transportation Research Record: Journal of the Transportation Research Board* **2121**, 55–62.

Spiess, H. 1987, 'A Maximum Likelihood Model For Estimating Origin-Destination Matrices', *Transportation Research Board* **21B**(5), 395–412.

Trépanier, M., Tranchant, N. and Chapleau, R. 2007, 'Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System', *Journal of Intelligent Transportation Systems* **11**(1), 1–14.

van Diggelen, F. 2007, 'GNSS accuracy: Lies, damn lies, and statistics', *GPS World* **18**(1), 26–32.

Van Zuylen, H. J. and Willumsen, L. G. 1980, 'The most likely trip matrix estimated from traffic counts', *Transportation Research Part B: Methodological* **14**(3), 281–293.

Wang, W., Attanucci, J. P. and Wilson, N. H. M. 2011, 'Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems', *Journal of Public Transportation* **14**(4), 131–150.

Zhao, J., Rahbee, a. and Wilson, N. 2007, 'Estimating a rail passenger trip origin-destination using automatic data collection systems', *Computer-Aided Civil and Infrastructure Engineering* **22**(5), 376–387.

Zhao, J., Zhang, F., Tu, L., Xu, C., Shen, D., Tian, C., Li, X. Y. and Li, Z. 2017, 'Estimation of Passenger Route Choice Pattern Using Smart Card Data for Complex Metro Systems', *IEEE Transactions on Intelligent Transportation Systems* **18**(4), 790–801.

Zhao, Z., Koutsopoulos, H. N. and Zhao, J. 2018, 'Individual mobility prediction using transit smart card data', *Transportation Research Part C: Emerging Technologies* **89**(August 2017), 19–34.