Available online at www.iujtl.com

Journal of Transportation and Logistics
7 (2) 2022

DOI: 10.26650/JTL.2022.1192128

RESEARCH ARTICLE

A Method for Public Transit OD Estimation Using Transit Smart Card Data

Akif Fidanoğlu[1], **Ilgin Gökaşar**[2]

ABSTRACT

With the utilization of Intelligent Transportation Systems (ITS) in public transportation (PT), trip data of passengers are recorded continuously by transit agencies. Researchers developed various algorithms to infer the origin and destination (OD) points of the PT users by using their transit smart card data. The origin and destination matrix is a very vital metric that can help the authorities to analyze public transportation. This study suggests a comprehensive method for OD inference of trips taken on public transportation. The OD pairs of the trips that cannot be inferred using chaining trips are found using the trip histories of the passengers. For the trips that have no further information, we used the overall OD distribution of the studied route.

Keywords: DC Systems, Intelligent Transportation Systems, Origin and Destination Matrix, Public Transit OD Estimation

Submitted: 20.10.2022 • **Accepted:** 03.01.2023

[1] **Corresponding author:** Akif Fidanoğlu (Dr.), Istanbul University, Faculty of Transportation and Logistics, Istanbul, Turkiye. E-mail: akif.fidanoglu@istanbul.edu.tr ORCID: 0000-0003-1227-4984

[2]Ilgin Gökaşar (Assoc. Prof.), Boğaziçi University, Department of Civil Engineering, Istanbul, Turkiye. E-mail: ilgin.gokasar@boun.edu.tr ORCID: 0000-0001-9896-9220

Citation: Fidanoglu, A., & Gokasar, I. (2022). A method for public transit OD estimation using transit smart card data. *Journal of Transportation and Logistics, 7*(2), 579-588. https://doi.org/10.26650/JTL.2022.1192128

1. Introduction

PT ridership has been traditionally measured by regularly counting the number of boarding and alighting passengers at each stop. Transit agencies periodically carry out surveys to validate and supplement these measurements (Gordon, Koutsopoulos, and Wilson, 2018). Transit agencies have accumulated a variety of transportation data that can be used for a variety of applications thanks to the growing use of Automated Data Collection (ADC) Systems. ADC systems have been shown to provide sufficient data for origin and destination (OD) inference. With the utilization of these technologies, the disadvantages of conventional data collection methods like small sample size, high cost, biased samples, etc., were eliminated.

The origin and destination matrix of public transportation users can be utilized in various ways. OD matrices can provide insight into the performance of the bus routes. The inferred OD matrices can be used to determine the number of passengers, maximum loads for the route, interchange locations, average interchange duration, and several other pieces of information. These metrics guide the transit agencies in service design, bus scheduling, route optimization, etc. Therefore, OD matrices provide valuable information for transit agencies to improve the service quality of public transportation systems.

To infer the origins of passengers, Automated Fare Collection (AFC) data and the Automated Vehicle Location (AVL) were used. The AFC and AVL data were matched by the time stamp to detect the boarding location of the passengers (Trépanier, Tranchant, and Chapleau, 2007; Cui, 2006). By defining several guidelines and recognizing the trips that adhered to these predefined rules, many researchers sought to deduce the destinations of the passengers in their studies. Barry et al. (2002) assumed that a rail trip ends at the stop that is closest to the boarding of the first trip of the day and that a trip's destination is the station that is closest to the start of the subsequent trip. Zhao et al. (2007) and Wang et al. (2011) used similar assumptions to infer the rail destinations and bus destinations, respectively. Munizaga and Palma (2012) utilized a generalized cost function that uses the disutility of time and distance to infer bus destinations. Sánchez-Martínez (2017) extends the generalized-cost approach to rail networks and uses dynamic programming to infer transfers on multileg journeys. Jung and Sohn (2017) used a supervised machine-learning model to infer the origins and destinations of bus passengers based on land-use characteristics and smartcard data.

In this study, we proposed a multistep process for OD pairs of public transportation trips. The studied routes are bus routes but trip data of the users acquired from other public transportation modes (metro, tram, etc.) are also used to infer the destination points. To improve the OD inference rate of the suggested algorithm, this study aims to make the most of the recorded trip data.

2. Case Study

The studied bus routes are operating near the university region. 8 different bus routes shown in Table 1 were analyzed. Routes of the studied bus routes are shown in Figure 1.

Table 1. The number of trips and GPS data of the studied bus routes.

Bus Route ID	Total Trips	Total GPS Data
4/G	447,571	648,561
35/G	424,058	658,869
1/T	285,155	580,732
3/G	239,009	670,962
E/12	156,194	570,695
E/13	132,307	574,391
B/33-K	98,673	343,780
43/D	78,824	419,730
Total	**1,861,791**	**4,532,927**

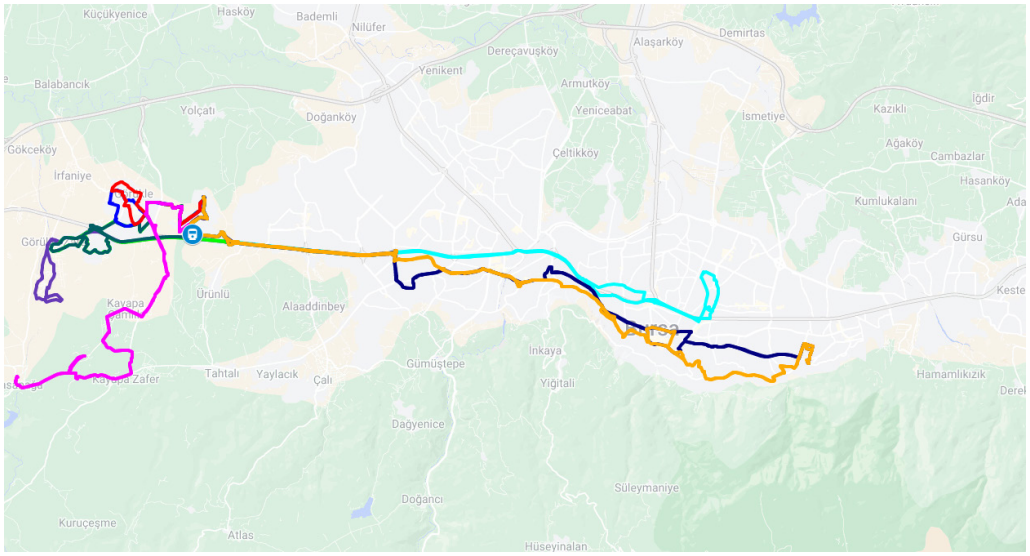

Figure 1. Routes of the studied bus routes.

Although the number of studied bus routes is low compared to the total number of bus routes operating in the city, searching for certain information like the trips made before and after the studied trips, was necessary. Therefore, over 180 million GPS data and 60 million trip data were analyzed. After making the required analysis of big data, trips made on the bus routes operating around the university area were obtained. 1.86 million trips were counted in total. Over 4.5 million GPS data were gathered from the AVL data of the buses that were utilized on the relevant bus routes.

2.1. Boarding Data

Boarding data analyzed in the study contains the three months of PT trips. These data were collected via Automated Fare Collection (AFC) Systems when the passengers use the smartcard while boarding the bus. The boarding data has the following information:

- Time and date of the trip
- The ID of the bus route
- Card ID of the passenger
- Vehicle ID

- Bus route ID of the previous trip
- Vehicle ID of the previous trip

2.2. GPS Data

The AVL system tracks the vehicles in the fleet dynamically. The locations of the vehicles are detected and recorded at a high frequency. Because of the need for precision in vehicle location data, the size of the GPS data reaches high numbers when the studied time becomes too long. Hence computational efforts for the operations made on GPS data might be considerably high when dealing with long periods. In our GPS data following information was available:

- Date and time of the data
- The ID of the bus route
- Vehicle ID
- Position of the vehicle by latitude and longitude

3. Methodology

In this section methods proposed in previous studies for origin and destination inference of bus passengers will be discussed. Additional assumptions that we utilized in this study for further OD inference are also shown.

3.1. Origin inference

In most of the studies, AFC data contains information on the boarding location of the passenger. This information can be achieved by matching the boarding data with the GPS data of the bus. The bus's GPS data is used to locate and search for the timestamp in the boarding data. The GPS data's nearest timestamp is chosen, and the position data for that timestamp is obtained.

However, in our study, origin inference required further analyses because of the data structure. The data acquired from the GPS data of the buses are the latitude and longitude of the buses. These GPS data needed to be converted to bus stops along the related bus route. To achieve this, the positions of the bus stops should be known. Then, the car was assigned its nearest bus stop based on its newly acquired position. This operation could be performed either for the entire vehicle's GPS data or just the timestamps in the boarding data. In our study, GPS data don't contain the direction of the buses. To infer the correct boarding stop of the passenger, it is necessary to know the direction of the bus. To detect the directions of the buses, all the GPS data of the buses were analyzed in our study.

It is not possible to determine the direction of the vehicle by a single location data. The previous and the following locations of the bus are also needed to infer the direction of the bus. From the sequence of the locations vehicle's movement can be detected. Due to this, GPS data was analyzed throughout the day of the study. The closest bus stops in both directions were identified for each GPS data. Then, the direction which gives

an increasing order was selected as the actual direction of the vehicle. This process was carried out for every vehicle that operated in the studied bus routes on the studied dates.

The boarding locations and the inferred destination locations are all based on the positions that were acquired from the GPS data. Therefore, even a small error in GPS data may result in an incorrect inference of the OD pairs. In our study, we have experienced various types of errors resulting from GPS data. Several methods were introduced to eliminate these errors. For simplicity, only one of them will be stated in this thesis.

The GPS data used in our study sometimes failed to give accurate locations of the vehicle. In some cases, the location data taken from the GPS data was wrong but considered close to the vehicle's actual location. This error can be detected after further analysis of these data points. The data points that were considered to be wrong were compared to the previous and next GPS data of the vehicle. The GPS data positions allow for easy assessment of the vehicle's movement. However, the increasing order and the progress of the vehicle are disrupted by these inaccurate data. If the algorithm doesn't have a way to fix these mistakes, it constructs a different route from the locations where the incorrect data refers. To tackle this problem, instead of getting the nearest bus stops to the GPS data, all bus stops within a certain radius were determined. This process produces not an exact bus stop but possible bus stops that can be assigned to the studied GPS data. A bus stop in the possible stops that does not break the vehicle's movement is selected and assigned to the related GPS data as the boarding stop.

3.2. Destination inference

To infer the OD pairs of transit users, the location information of the vehicles must be known. For this reason, transit agencies all around the world equipped their vehicles with Global Positioning Systems. The location data of the vehicles and the boarding time of the users are matched. Hence the origin locations of the passengers can be inferred. However, most cities don't have transportation systems that record the alighting stops of the users (Jung and Sohn, 2017). In most PT systems, passengers only use their smart cards when boarding. Therefore, there is a need for further inference study to get the alighting locations of the users.

Previous studies made assumptions to infer the destinations of the passengers (Zhao, Rahbee, and Wilson, 2007; Cui, 2006; Trépanier, Tranchant, and Chapleau, 2007; Barry et al., 2002; Wang, Attanucci, and Wilson, 2011). These assumptions are similar to each other and can be summarized as follows:

- No other transportation mode is used between the recorded PT trips.
- Passengers do not walk between the PT routes above a predefined distance.
- The boarding location of the first trip of the day is assumed to be the destination of the last trip of the day if it doesn't exceed walking distance limitations.

Fidanoglu (2015) further proposed the assumption that if there is a missing leg in the chained trips of the passenger that can be easily determined, the destination inference

should be made from the missing trip. This assumption is valid for some instances where the passenger cannot physically make the recorded trips without introducing the missing leg. To detect this the missing leg of the trips should be obvious. However, in our study, we did not assume any missing trips besides the recorded trip.

To determine the destination of a trip, all relevant trip data of the same passenger should be determined. In this study, the following trips of the passengers were also studied:

- Previous trip of the passenger
- Next trip of the passenger
- First trip of the passenger on the day of the studied trip
- Last trip of the passenger on the day of the studied trip.

All of the above trips serve for further inferences of destination. For some studied trips, the above trips cannot be obtained due to the lack of information. It is also possible to acquire such trips that are irrelevant to studied trips. For example, a subsequent trip made after five days of the studied trip doesn't give any direct information about the trip's destination.

The destination inferences based on the next trips are the most reliable method if the next trip is made within a short time. Therefore, at the first stage of the destination inferences, the trips made within 2 hours after the studied trip were analyzed. If the boarding location of the next trip is close enough to any bus stop on the route of the studied bus route, the nearest stop is taken as the destination of the trip. The maximum distance was limited to approximately 500 meters for this method.

Next, inferences from the first trip of the day were studied. If the trip is the last trip of the day, it is checked whether the passenger has a different trip that is the first trip of the passenger on that day. If the first trip is different from the last trip passenger, meaning that the passenger made at least two trips on the studied day, the boarding location of the first trip is determined. Then, the distance between the boarding location and the closest bus stop of the studied bus route is calculated. The nearest bus stop is assumed to be the final stop of the day if the distance satisfies the requirements. About 300 meters was chosen as the method's maximum range. The main logic behind these inferences was that at the end of the day, passengers return to the locations where they started their trips on that day. Therefore, the maximum distance was lower compared to the distance used in the previous destination inference method.

If the inferences from the interchange within 2 hours or the first trip of the day could not be achieved, the next trips made within 24 hours after the studied trip are determined. The same procedure used in the first destination inference method is followed for these trips also.

Next, the proposed algorithm aims to infer destinations for the first trips of the day from the last trips of the day. If the boarding of the last trip is within a distance of 300 meters of

any stop in the studied route, the closest stop to the boarding of the last trip is assigned as the destination of the first trip. In fact, the previous method includes destination inference from the last trip of the day if the next trip after the first trip of the day is the last trip of the day. However, in some cases, there are some other trips between the first and the last trips of the day. For some of those cases, destination inference cannot be achieved from the interchanges. Therefore, this step is included in the proposed method.

For all of the destination inference methods, a stop correction method was introduced. This method aims to correct the inferred destination stop if a direction error occurs. For the bus routes that follow a circular route, some stops are very close to each other. When the closest stop to a certain location is determined for destination inference, it is highly possible to select the wrong stop. In this case, a direction error will rise. To eliminate this problem, the introduced method searches the stops nearby and assigns the other bus stop if it resolves the direction problem.

In some of the proposed inference methods, the time between the studied trip and the trips that the destination inference was made from can be considered too long. However, for PT users, this kind of trip pattern is not something unexpected. In addition, even if the time difference between the trips is too long, estimating the destination of the trip using other methods would require further assumptions. Therefore, it is assumed that if the passenger can be tracked within the routes of his trips, the destination inference is valid.

All of the above-mentioned methods that were used for the destination inference of the trips check the direction error. If the inferred boarding stop is behind the boarding stop of the studied trip in the direction of the bus, it is concluded that there is a direction error. In these cases, inferred destinations were not accepted.

3.3. Derived and Assigned Destinations

At the end of destination inferences, some of the trips were still missing destinations. To infer destinations for these trips, inferred OD pairs were used. The OD matrix contains the ID for each cardholder. Therefore, it is possible to obtain the inferred OD pairs for each passenger. A data frame including the following information was generated:

- The ID of the bus route
- Card ID
- The direction of the bus route
- Inferred boarding stop
- Inferred alighting stop
- Number of trips

For every origin point, the probabilities of destination points were determined by the number of trips. For a trip missing a destination, the passenger Card ID is determined, and the above data frame is constructed. If the data frame contains the direction and the

boarding stop of the studied trip, meaning that the passenger made a trip similar to the studied trip, the alighting stop is probabilistically inferred.

If the inference was not achieved from other trips of the cardholder, the overall distribution of OD pairs was used for destination inference. To achieve this, inferred OD pairs for the studied bus route were clustered into three groups of time zones. Time intervals of the categories were 06:00-10:59, 11:00-15:59, and 16:00-23:59. This categorization was done because the patterns of the trips during the morning and evening hours were different.

Next, the time of the studied trip and the boarding stop was determined. From the inferred OD pairs, trips sharing the same time zone, direction, and boarding stop with the studied trip were extracted. A destination point from these pairs was probabilistically selected and assigned to the studied trip.

Previous studies evaluate the performance of the proposed algorithm by the inference rate. However, in our research, if the algorithm fails to infer a destination for a passenger from his or her trips, it searches for similar trips through the inferred OD pairs and assigns a destination to the trip. Therefore, at the end of the destination inference process, all trips have inferred destinations.

4. Results

Inference algorithms are coded in Python 3.9. The inference method used in this study has certain differences from the trip-chaining methods used in the previous studies. The inference method in this study gives very successful results. The proposed methods failed to infer OD pairs for very few trips.

Table 2. Inferences of bus routes

Bus Route ID	Inferred	Derived	Assigned	No results	Total
4/G	311.554	83.493	52.450	74	447.571
35/G	303.250	75.478	45.226	104	424.058
1/T	215.009	45.718	24.320	108	285.155
3/G	170.896	34.538	33.551	24	239.009
E/12	114.092	14.580	27.471	51	156.194
E/13	95.208	13.502	23.516	81	132.307
B/33-K	72758	13640	12160	115	8.673
43/D	54.440	11.124	13.178	82	8.824

As seen in Figure 2 inference rate of the proposed algorithm is around 70%. These OD pairs are inferred by using chaining trips of the passengers. The most reliable source for inference is the interchanges made within a short period. Moreover, the first and the last trips of the passenger on the studied trip are also used for further inference. All these inferences are plotted in Figure 2 as “Inferred”.

For additional inference, we used the trip history of the passenger on that specific route. If there are inferred OD pairs of the studied passenger in the studied route algorithm search for a similar pattern in those OD pairs. Then a destination point is assigned to the studied trip probabilistically. This means that if the passenger has a different OD pattern

in his or her trip history it is possible to have a different inferred destination in this stage of the algorithm. This step of the algorithm is plotted in Figure 2 as "Derived" since the destination is derived from the trip history. As seen in Figure 2 an average of 15% additional inference is achieved by using this method.

Lastly, the overall OD distribution acquired from the earlier steps of the proposed algorithm is used for the destination assignment process. This is shown in Figure 2 as "Assigned". OD distribution of the studied route is grouped into different time zones of the day. Then the trips that the previous step of the algorithm couldn't achieve to assign destination points are analyzed. Destination points are assigned to those trips by using the time zone of the trips and the OD distributions. This stage is also carried out probabilistically. Around 15% extra OD inference is achieved by the utilization of this step.

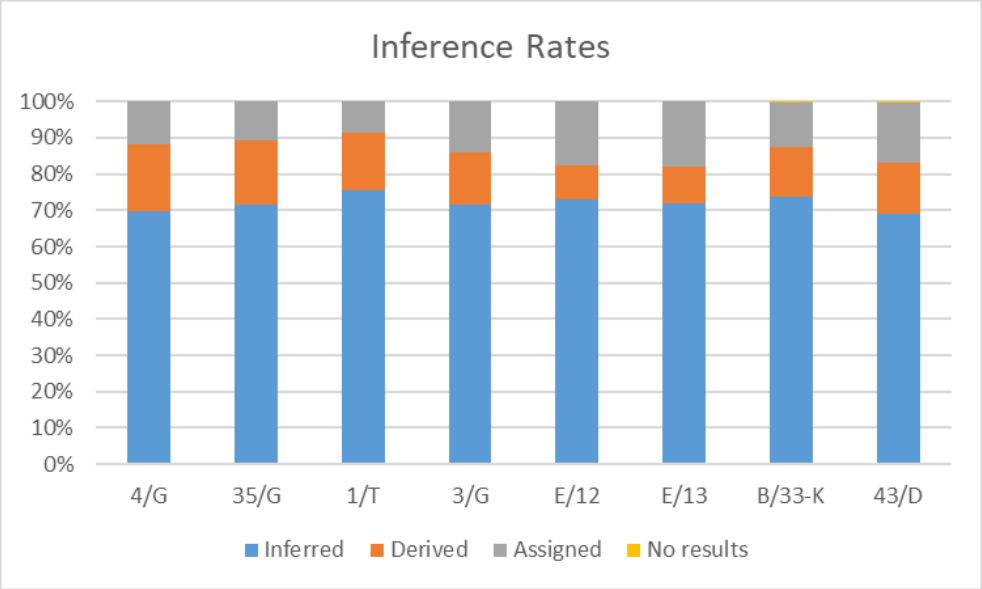

Figure 2. Inference rates bus routes.

5. Conclusion

ADC systems offer a rich database to the transit agencies which can be utilized for public transportation planning. Data acquired from smart card transactions yield valuable information on trip demands and trends. This data can also be used for the inference of OD pairs in public transportation. The estimation of OD pairs is highly critical for public transportation planning and management. This study proposes a multistep framework for inference of OD pairs of public transportation trips. The proposed model used passengers' trip chains by several assumptions.

The results show that for over 70% of the trips OD pairs can be inferred by the information from chained trips of the passenger. This study further analyzed the trip data and aimed to increase the inference rate by using the passengers' other trips and inferred OD pairs. This results in approximately 15% further OD inference. Other portions of the trip data that have no inference from the previous methods are also analyzed. Overall, the OD distribution of the studied route is used to assign destination

points to those trips. To achieve that trips are categorized based on their time zone for more precise estimation.

The proposed method for OD inference shows its capabilities in the estimation of OD pairs of public transportation trips. The real-world data taken from the transit agency was analyzed and applied the proposed method to that data. In this sense, all the difficulties and problems related to real-world applications needed to be solved for this study.

Peer Review: Externally peer-reviewed.
Conflict of Interest: Author declared no conflict of interest.
Financial Disclosure: Author declared no financial support
Author Contributions: Conception/Design of Study- A.F., I.G.; Data Acquisition- A.F.; Data Analysis/Interpretation- A.F., I.G.; Drafting Manuscript- A.F., I.G.; Critical Revision of Manuscript- A.F., I.G.; Final Approval and Accountability- A.F., I.G.
Acknowledgement: This work was made possible through the support of BURULAŞ, the public transit agency of the city of Bursa, Turkey

References

Barry, J.J., R. Newhouser, A. Rahbee, and S. Sayeda, 2002, "Origin and Destination Estimation in New York City with Automated Fare System Data." *Transportation Research Record* 1817 (1): 183–87.

Cui, A., 2006, *Bus Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems*. M.Sc. Thesis, Massachusetts Institute of Technology.

Fidanoglu, A., 2015, *Origin and Destination Inference of Bus Passengers: Istanbul Case Study*. M.Sc. Thesis, Boğaziçi University.

Gordon, J.B., H.N. Koutsopoulos, and N.H.M. Wilson, 2018, "Estimation of Population Origin–Interchange–Destination Flows on Multimodal Transit Networks." *Transportation Research Part C: Emerging Technologies* 90 (May): 350–65.

Jung, J., and K. Sohn, 2017, "Deep-learning Architecture to Forecast Destinations of Bus Passengers from Entry-only Smart-card Data." *IET Intelligent Transport Systems* 11 (6): 334–39.

Munizaga, M.A., and C. Palma, 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 (October): 9–18.

Sánchez-Martínez, G.E., 2017, "Inference of Public Transportation Trip Destinations by Using Fare Transaction and Vehicle Location Data: Dynamic Programming Approach." *Transportation Research Record* 2652 (January): 1–7.

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

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

Zhao, J., A. Rahbee, and N.H.M. Wilson, 2007, "Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems." *Computer-Aided Civil and Infrastructure Engineering* 22 (5): 376–87.