Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems

Wei Wang, The World Bank Group

John P. Attanucci and Nigel H.M. Wilson
Massachusetts Institute of Technology

Abstract

This research explores the application of archived data from Automated Data Collection Systems (ADCS) to transport planning with a focus on bus passenger travel behavior, including Origin-Destination (OD) inference, using London as a case study. It demonstrates the feasibility and ease of applying trip-chaining to infer bus passenger OD from smart card transactions and Automatic Vehicle Location (AVL) data and is the first known attempt to validate the results by comparing them with manual passenger survey data. With the inferred OD matrices, the variations of weekday and weekend bus route OD patterns are examined for planning purposes. Moreover, based on the inferred OD matrices and the AVL data, alighting times for bus passengers also can be estimated. Bus journey stages, therefore, can easily be linked. By comparing the interchange time and the connecting bus route's headway, it provides a way to evaluate bus connections.

Background and Purpose

London has one of the largest bus networks in the world, with more than 6 million passengers transported on its 700 routes daily. A recent report states that "Bus usage is growing at its fastest rate since 1946. More than two billion passenger trips were made on London's fleet of more than 8,000 buses in the year to March 2009. The number of operated kilometers has also risen to 478 million, the highest since 1957" (Mayor of London, 2009).

Every five to seven years, a Bus passenger Origin and Destination Survey (BODS) is conducted by Transport for London (TfL) for each bus route. This survey provides detailed information about passenger travel patterns, including the number of people boarding and alighting at each stop, the purpose of travel, the boarding and alighting locations for each journey, and how passengers get to the boarding stop and from the alighting stop to their final destination. Expansion factors are used to account for non-returned survey cards and non-surveyed bus trips. An automated database (BODS database) stores survey results, including boardings, alightings, and loads at each stop (or stop zone) for each route. BODS is one of the primary data systems used by the Bus Network Development Unit at TfL. A limitation of this type of survey is that it records passenger travel for only one day per route. Recognizing the substantial network growth and the dynamics of demand, supplementary data from other sources also are needed for network planning. Moreover, although surveyed passengers are asked for their ultimate origin and destination in addition to their travel on the route itself, this information is not generally transferred from the paper surveys into the BODS database and therefore is not readily available to network planners.

In addition to BODS survey data, London bus planners also get timely route-level passenger ridership data from Electronic Ticketing Machine (ETM) transactions, which are downloaded daily from each bus at the garage. One drawback of this data collection method is that this data source only records aggregate ridership for each bus trip, while detailed information such as boarding and alighting locations for each passenger cannot be obtained directly.

The Oyster smart card system was launched in London by TfL in December 2003 as a new ticketing medium (Transport for London). It is now accepted on the Underground, buses, the Docklands Light Rail (DLR), Tramlink, and National Rail stations. Though the full potential of this data source has not yet been realized by London bus network planners, Oyster data are readily available, provide large sample sizes, and potentially offer a full network perspective rather than strictly a

mode level view. Bagchi and White (2004) summarize the benefits of smart card data systems as follows: (1) much larger volumes of individual passenger trip data than from manual surveys; (2) the potential to link individual passenger trips to individual cards or travelers; (3) continuous trip data covering longer time periods than manual surveys, allowing for panel data analysis techniques; and (4) classification of different customer market segments using transit services.

In addition, using Oyster smart card data enables one to link trip segments and to determine OD flows across the network. This process can be repeated on a daily basis to assess variability in trips and get more accurate estimates of ridership for specific days of the week and times of the year. It provides an easier and more reliable way to get more detailed passenger behavior information than manual survey data, which potentially can help transit agencies improve efficiency and reduce cost.

Literature Review

Cui (2006) summarized OD estimation techniques using manually-collected data. Basically, the OD matrix can be obtained either from surveys or through techniques that combine various sources of data. The ever-increasing use of ADCS generates new transport data that can be used by service providers for a range of applications. Although most ADCS are designed to support specific agency functions, the resulting data can be applied to areas far beyond their design purposes. Recent research has examined the potential benefits of using ADCS for public transport planning, specifically using archived ADCS data to infer OD matrices to assess service performance for service planning. Because most Automatic Fare Collection (AFC) systems record the bus trip boarding location coarsely at the bus-route level, it is still difficult to obtain information about where individual passengers board a bus. Integration of the AFC system data, which includes characteristics of each fare card transaction, with the AVL system data, which includes vehicle locations, offers a solution through matching the vehicle location information with the passenger trip information to help transit planners infer individual passenger boarding locations (Cui 2006).

To infer the destinations for individual passenger trips, Zhao et al. (2007), Cui (2006), Trepanier et al. (2007), and Barry et al. (2008) all used trip-chaining methods with assumptions similar to those summarized by Zhao et al. (2007):

- There is no private transportation mode trip segment (car, motorcycle, bicycle, etc) between consecutive transit trip segments in a daily trip sequence.

- Passengers will not walk a long distance to board at a rail/bus station different from the one where they previously alighted.
- Passengers end their last trip of the day at the station where they began their first trip of the day.

Jang (2010) further examined the possibilities of using the ADCS archived data for public transport planning in Seoul, South Korea. One feature that distinguishes the Seoul ADCS from many other cities is that it records each trip's entry and exit times and locations, as well as the trip chains with interchanges. Based on this dataset, Jang analyzed interchange patterns and identified interchange points that needed improvement by examining the points where interchange demand exceeded 5,000 per day and/or the average interchange time exceeded 10 minutes.

Method Applied in London

The transit passenger OD estimation methodology used in this research builds upon the trip-chaining OD estimation method applied in Chicago by Cui (2006). Since different transit agencies may have different data sources with different characteristics, the next step is to describe the data sources used in the London application.

TfL ADCS Introduction

Oyster Smart Card Data

Oyster is the contactless smart card used for public transport for fare payment in London. It has a penetration rate of around 85 percent for all bus passengers in London. Oyster smart cards in London are owned by individuals and record every transaction the card holder makes while traveling on the public transportation system. For the Underground and Overground networks, generally both the time and rail station for entry and exit are recorded. However, for buses, only the time of passenger boarding and route number are recorded. Several types of analyses are possible with the smart card data, including ridership monitoring, revenue estimation, and service performance measurement. The key contribution of this research, however, is to develop a methodology to infer the origins and destinations for bus passengers in London using the Oyster data and to develop related applications for the London bus network.

iBus Data

iBus is a £117m AVL and radio system that aims to help London Bus Services Limited run more reliable and consistent bus service (Hardy 2009). The first installa-

tions took place in March 2007, with system-wide deployment completed in April 2009. iBus data contain information about the route and trip number as well as the direction for each bus trip, and most important, they provide a unique bus stop identifier and record the departure time from each stop.

Methodology Based on Oyster and iBus Data

The basic premise is that it should be possible to determine the boarding stop for every passenger who uses an Oyster card to board an iBus-equipped bus. For a given route and trip, the fare collection timestamp (including the date) from the Oyster card is used to search through the iBus dataset to determine the boarding stop and vehicle ID. The boarding location of the next trip taken by the passenger is then used to infer the alighting stop, where possible.

Origin Inference

Since the Oyster system records only the timestamp when an Oyster card user boards a specific bus, but no location information, while the iBus AVL system records the time when the bus doors open or close at each bus stop for each bus run, it is possible to determine the boarding stop by matching the Oyster transaction times with the corresponding iBus data. In this case, the origin inference procedure is implemented through a custom-built Java program.

Destination Inference

The destination inferences are based on the trip-chaining method and use the same assumptions proposed by Zhao et al. (2007), Cui (2006), Trepanier et al. (2007), and Barry et al. (2008), as described above. The destination inference is implemented in a custom-designed Java program that reads its inputs from an SQL database.

The procedures to implement this methodology are illustrated in Figure 1. The process begins by checking whether the bus fare transaction under examination is the only Oyster transaction for that card on that day. If it is, then the trip-chaining method cannot be applied and, thus, the trip destination cannot be inferred. Otherwise, it is determined whether this bus fare transaction is the last of the day for this card. If it is not, the trip-chaining method is applied by 1) determining whether the next fare transaction for this card is on bus or rail; 2) if the next transaction is also on bus, the algorithm moves onto the "next trip" rule with a bus lookup table sub-procedure; 3) if the next transaction is on rail, the algorithm moves onto the "next trip" rule with a rail lookup table sub-procedure. If the fare transaction currently under examination is the last of the day for that card, then the first transaction of the day is treated as the transaction immediately following this last trip

segment so that the "next trip" rule can be applied here to infer the destination of this last trip segment.

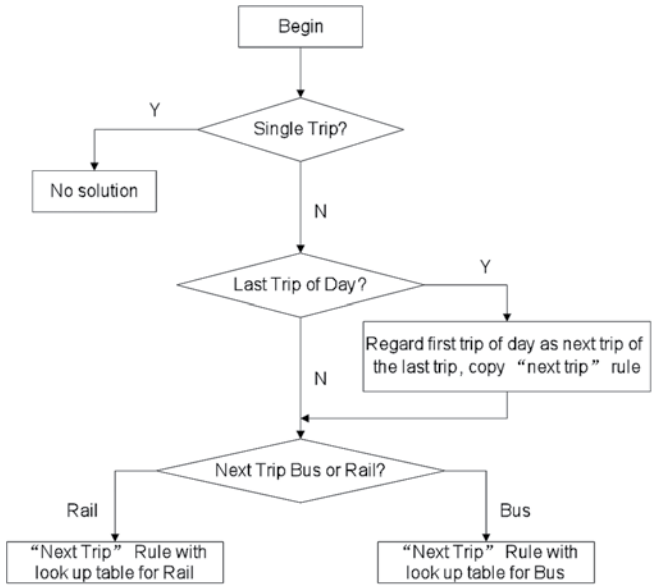


Figure 1. Process for destination inference

The lookup table mentioned here defines the stops on the bus route under examination that are closest to the boarding stop of the next transaction. While the two sub-procedures for bus and rail are similar, the London rail and bus networks are in two different GIS files, and the lookup tables are generated separately. The "next trip" rule is actually the same as assumption 2) listed in the literature review, meaning that travelers start their next trip segment at another station in close proximity (within walking distance, for example at most 1 km, or 12 minutes' walking distance at a speed of 5 km/h) to the destination of their initial trip segment.

OD Inference Results

Five routes in the London bus network are selected to test the OD inference procedures, including two connecting suburban areas, two that terminate in Central London and one thast runs through Central London. The results are shown in Table 1.

In general, the inference process has been shown to work fairly well. As shown in Table 1, origins can be inferred for more than 90 percent of all the bus passenger trips using Oyster cards on the five selected routes, and more than 57 percent of these bus passenger trips have both origins and destinations inferred. Such com-

prehensive information on a majority of bus passengers can provide very useful statistics on the use of service in complex transit networks.

Table 1. Origin and Destination Inference Results

Bus Routes	No. of Oyster Transactions	No. of Origins Inferred	% of Origins Inferred	No. of Destinations Inferred	% of Destinations Inferred
W4	8,585	8,212	95.7%	5,393	62.8%
70	12,074	11,381	94.3%	7,741	64.1%
185	2,4245	22,794	94.0%	13,947	57.5%
307	10,057	9,456	94.0%	6,968	69.3%
329	17,496	17,033	97.4%	13,737	78.5%

Validation

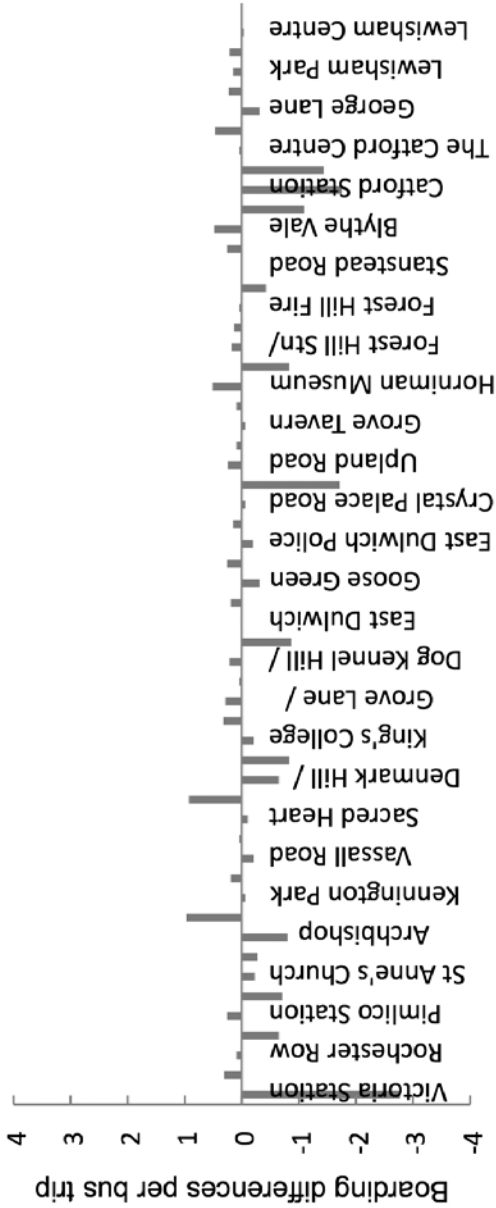
The next step was to validate the inferred origins and destinations for the selected bus routes in London. First, the origins inferred from Oyster transactions are compared with the BODS survey results for all the manually-surveyed bus trips. Then, the BODS surveyed destinations are compared with the results from the Oyster inference methodology for the same bus trips.

Comparison of Boardings for the BODS and Oyster Datasets

Since the origin inference rates are quite high and BODS does not survey every bus passenger (the sample rates for some bus trips are as low as 60 percent), the total number of boardings inferred from the Oyster transactions is close to that for the BODS survey. Table 2 summarizes the total number of boardings from BODS and Oyster datasets in terms of direction for the surveyed bus trips on Route 185. It shows that though Route 185 is one of the busiest bus routes running through Central London with a daily ridership of around 26,000, the number of boardings from BODS and Oyster for all the surveyed bus trips is quite consistent. Consequently, the number of boardings at each stop inferred from the Oyster transaction dataset should be close to that from the BODS database if the origin inference method works well. Figure 2 demonstrates the boarding location comparison results for this route.

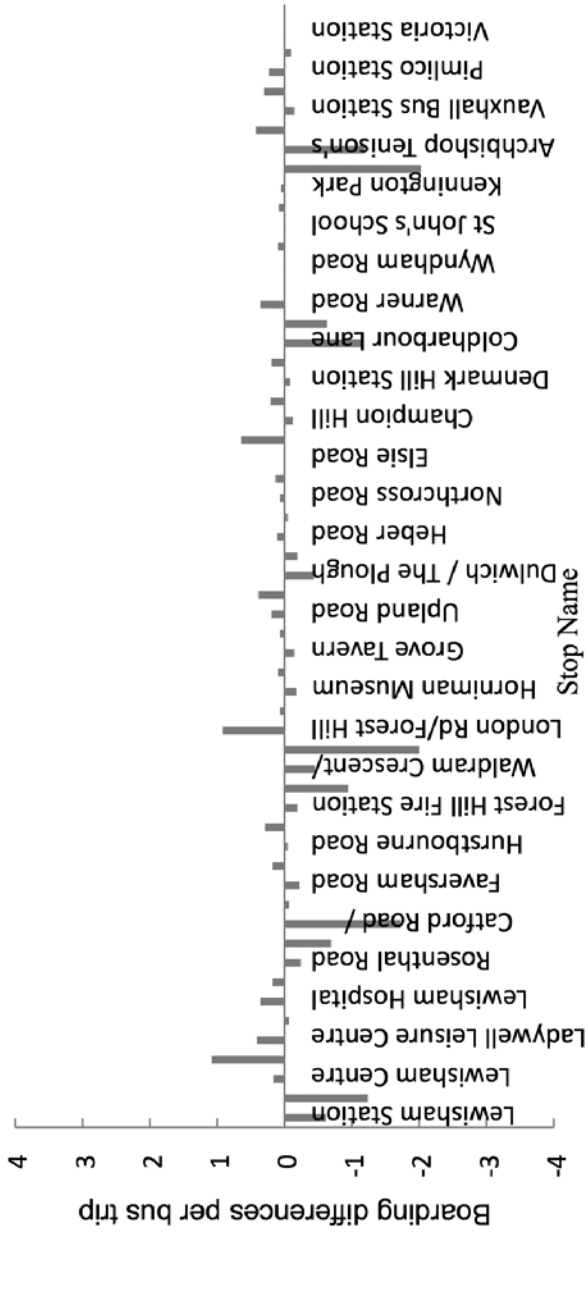
Table 2. Number of Boardings from BODS and Oyster (Route 185)

Direction	No. of BODS Boardings	No. of Oyster Boardings	No. of Surveyed Bus Trips
Eastbound	7,304	7,911	66
Westbound	6,904	7,386	62



(a) Eastbound

Figure 2a. Boarding locations for Route 185



(b) (Westbound)

Figure 2b. Boarding locations for Route 185

The total number of boardings for the surveyed trips from BODS is 607 fewer than that recorded in the Oyster dataset on Route 185 eastbound, and 482 passengers fewer westbound. These relatively large boarding differences between BODS and Oyster datasets are mainly due to the low BODS survey sample rates, as Route 185 is one of the busiest routes in London. Even so, the average difference per bus trip is small. As shown in Figure 2, most of the stops where the boarding differences are larger than one passenger per trip are close to shopping centers or stops with survey problems that are listed in the BODS summary report from TfL.

In general, the number of boardings at each stop from Oyster estimates is very close to that from the BODS survey. Some minor differences are caused by the low BODS sample rate. Overall, these and similar results from other routes studied show that the origin inference methodology works well and thus could be used to further infer bus passengers' destinations as well as to provide more comprehensive and reliable information for transit planners.

Comparison of Alighting Locations between BODS and Oyster Datasets

This section tests the destination inference methodology by comparing the percentages of alightings at each stop in the BODS dataset with the Oyster estimates. Since destinations could be inferred for only about 60 percent (see Table 1) of all the Oyster transactions on the studied bus routes, the number of inferred alightings at each stop from Oyster typically will be far less than the BODS survey result on any given bus trip. But it is expected that the percentages of inferred alightings from Oyster will be close to the percentages of alightings from the BODS dataset. Figure 3 demonstrates these results by comparing the alighting locations, again using Route 185 as an example.

As shown in the above figures, 7,386 alightings were recorded in the BODS survey, while destinations were inferred for 4,844 Oyster passengers (66% of BODS surveyed alightings) on Route 185 northbound. Southbound, destinations were inferred for 4,776 Oyster passengers (65% of BODS surveyed alightings). For both directions, the number of inferred Oyster alightings is far lower than the BODS survey results. However, the percentage of inferred alightings from Oyster at each stop is very close to the percentage of alightings from the BODS dataset, with the differences generally within two percent per bus trip. There is a relatively large difference (4%) between the BODS dataset and the Oyster estimates at the Catford Shopping Center bus stop southbound of Route 185. The BODS validation report provided by the BODS survey group in TfL mentioned some problems here, as several issued cards were not returned, which contributed to the difference. The other reason is

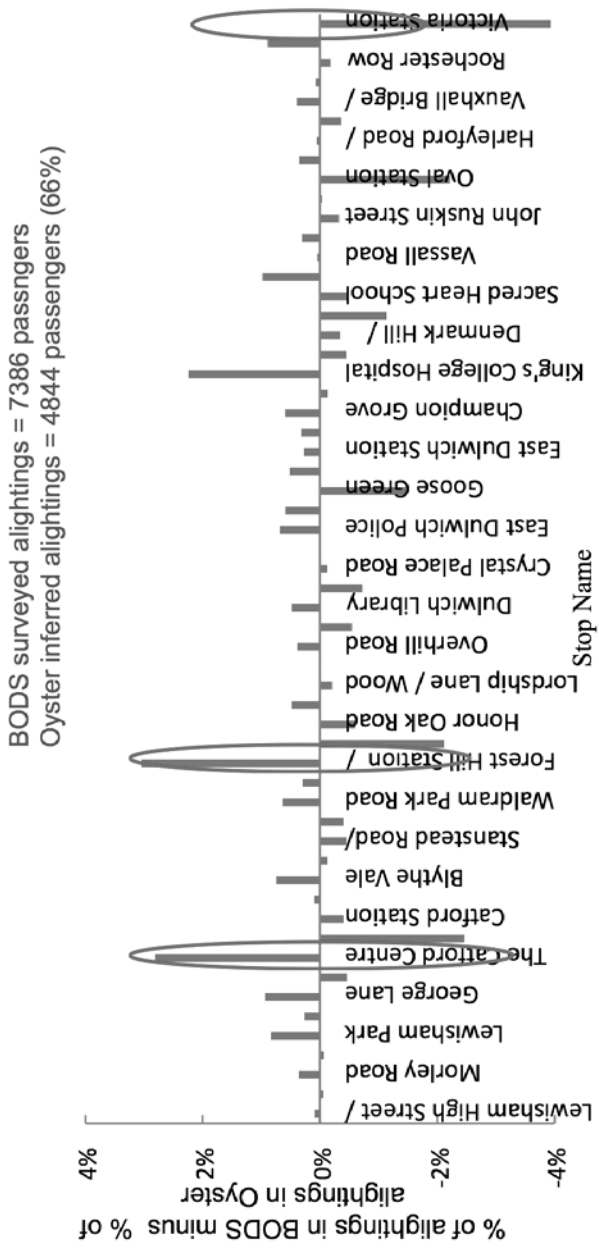


Figure 3a. Alighting location comparison for Route 185

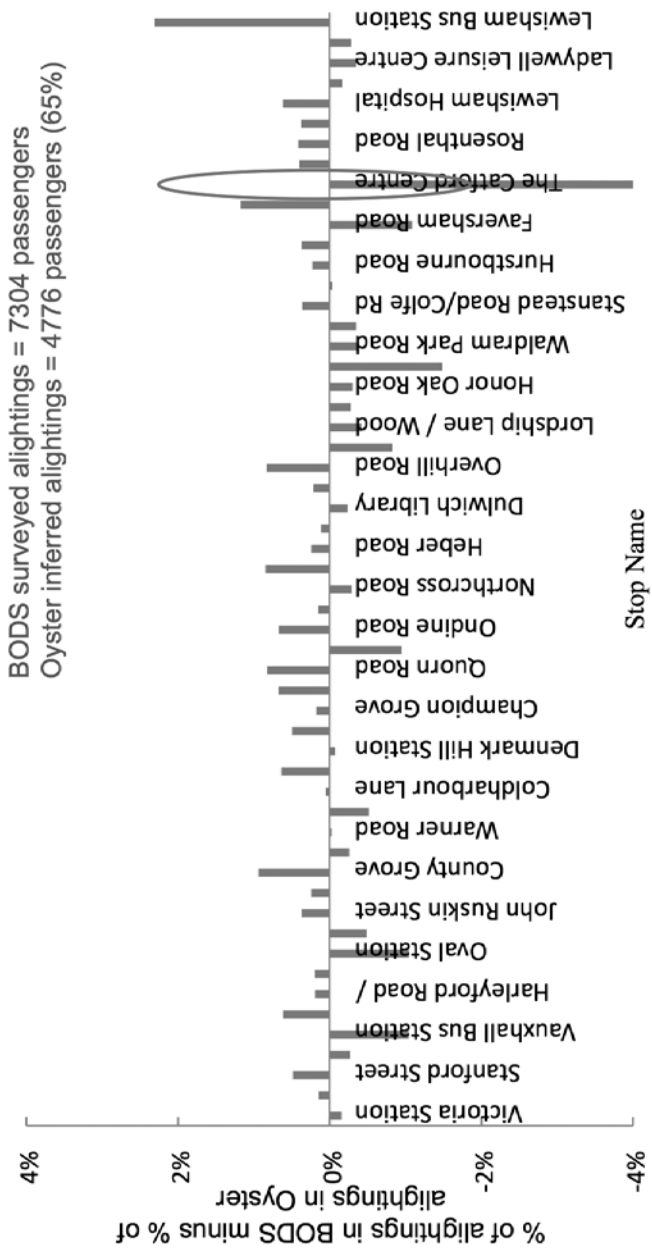


Figure 3b. Alighting location comparison for Route 185

that passengers might not necessarily get off the bus at the stop that is closest to their next boarding stop, especially when the stops are close to a shopping center, where people may walk further than usual. Another large difference appears at Victoria Station on Route 185 northbound, which is a major interchange hub and has connections with five other bus routes, as well as the Underground and National Rail. It is quite possible that the BODS survey did not reach all the passengers at this bus stop due to passenger crowding. However, for most of the other stops, if the percentage of alightings from the BODS dataset differs greatly from that from the Oyster estimates, these differences are generally offset by the differences at adjacent stops, as shown by the red circles in Figure 3. As mentioned above, passengers might get off the bus one or two stops away from their next boarding stop in order to walk and complete errands which our model cannot capture.

Application to Bus Network Planning

The validation has shown that the origin and destination inference process using the proposed methodology works well when compared with the BODS manual survey results. This section presents several applications using the results from the inference process. One of the most significant applications is using the automatic OD inference to better understand bus passenger travel patterns on a daily basis. Manual surveys are limited by narrow spatial and temporal coverage, while an automatic procedure can generate OD matrices for any bus route at any time at low marginal cost, as long as the ADCS and inference procedures have been developed and deployed.

Daily Load/Flow Profile Variation

Load/flow profiles are standard graphics showing passenger activity (boardings, alightings) and passenger load (or flow past a stop or segment in the case of multiple trips) along a route by direction. They allow planners to identify locations and values of the peak load, as well as underutilized route segments.

Route W4 during the AM Peak (7:00 to 9:30 AM) is chosen here as an example of how the daily load/flow profile varies over five successive weekdays. Figure 4 shows that there are large variations in the load/flow profile and specifically in the peak loads, even within the same week.

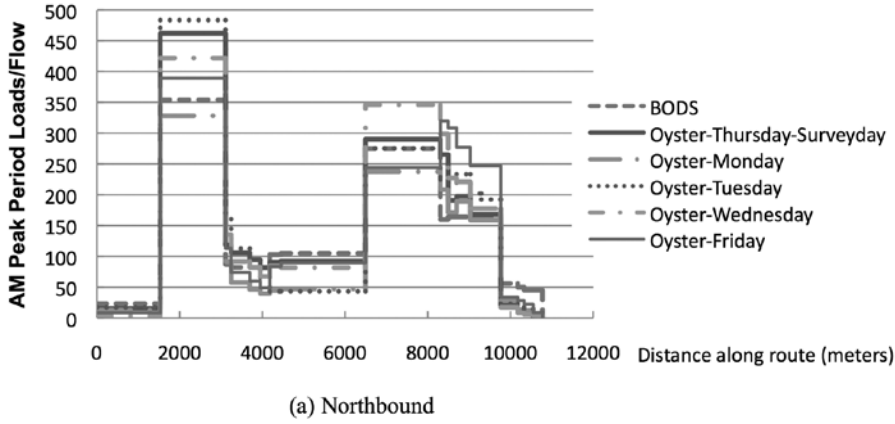


Figure 4a. Daily load/flow variation along Route W4 during AM peak

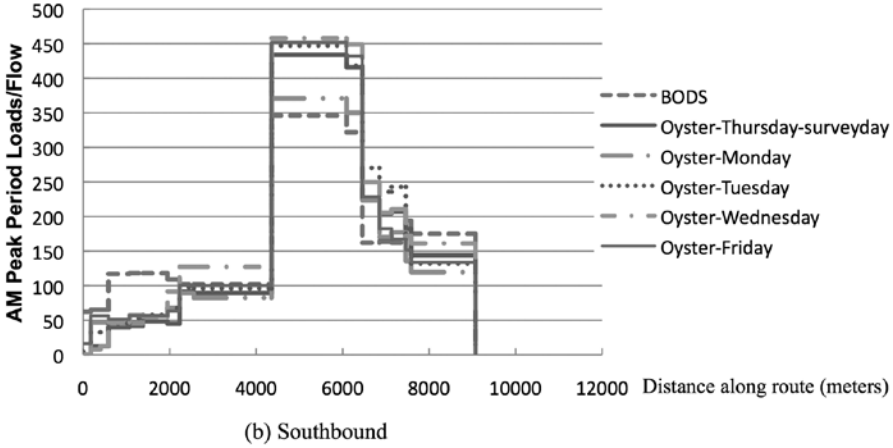
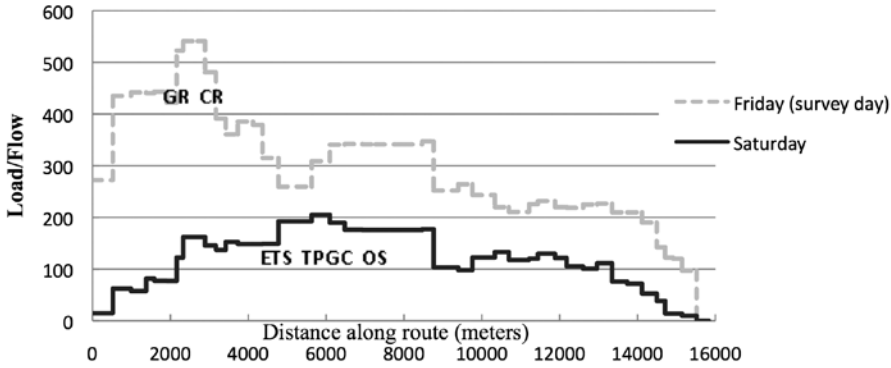


Figure 4b. Daily load/flow variation along Route W4 during AM peak

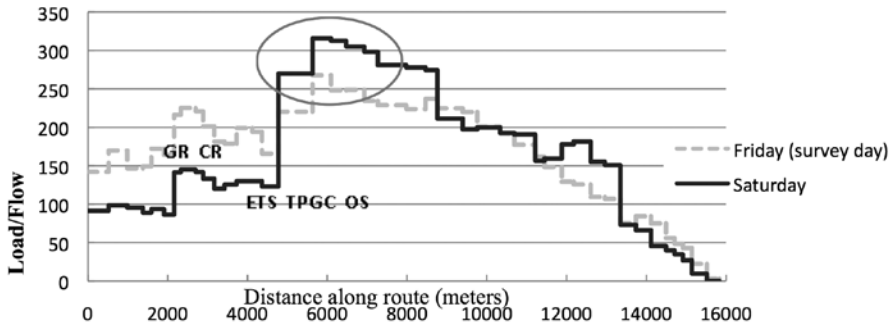
Since the OD information can be obtained for every day from the ADCS archived data, the load/flow profile differences for weekdays and weekends also can be studied. Figure 5 demonstrates the load/flow profile variations for a Friday and Saturday on Route 307. Generally, the load on Saturday is much lower than that on Friday, and the peak load point changes in the AM peak. On Friday, the peak load point is between Glyn Road (GR) and Crown Road (CR) while on Saturday, the peak load point is between Enfield Town Station (ETS), Trent Park Golf Course (TPGC)

and Oakwood Station (OS). It is quite likely that more passengers will visit the Golf Course on Saturday, which makes the load/flow around this bus stop higher than at other bus stops. In the PM peak (16 to 18:30), as shown by the circle in Figure 5(b), the peak load points are also around the Trent Park Golf Course bus stop and the Oakwood Station, but the loads/flows around these stops are even larger than on Friday. It is also likely that more people may transfer at Oakwood Station to other routes or the Underground on weekends for non-work trip purposes.



(a) AM Peak (Westbound)

Figure 5a. Daily load/flow profiles for Route 307



(b) PM Peak (Westbound)

Figure 5b. Daily load/flow profiles for Route 307

Interchange Time Analysis

Interchanges affect the attractiveness of public transportation and making interchanges less burdensome is a critical consideration in public transport planning. Improving the level of service at interchange locations would enhance the overall quality of public transportation services. Both practitioners and researchers tend to pay most attention to the initial waiting experience and to in-vehicle travel for their obvious effects on ridership, but less work has been done on interchanges between segments of a linked journey (Guo and Wilson, 2010). However, reducing the out-of-vehicle times can help make public transit more attractive resulting in ridership increases. In this research, bus passenger alighting locations are inferred from the ADCS archived data. Also, since iBus AVL data provide information about the observed departure time for each bus trip at each stop, by matching the inferred alighting locations with the iBus AVL data, the alighting time for each individual passenger trip can be estimated. Hence, the interchange time can be calculated more accurately as the difference between the subsequent trip's boarding time and the previous trip's alighting time.

Taking Route 185 as an example, based on the Oyster transactions, Route 176 is found to be the most frequently used connecting route for passengers originating from Route 185, with 15 interchange stops for the parallel segments. The median interchange time for passengers from Route 185 to Route 176 is five minutes. The most frequently used connecting stop on Route 176 is the Forest Hill Station (a stop shared by these routes, so interchange times for transfers at this stop do not include walking time, and thus are actual waiting times), with seven minutes as the median interchange time for passengers originating from Route 185. Transit planners often use half the headway of the connecting bus route as the estimated waiting time, but there are no field data to support such theory. The analysis in this research supports this assumption that the actual waiting time is approximately half the headway of the connecting route, as shown in Figure 6. In this figure, the size of each dot indicates the number of interchange passengers and the color indicates the scheduled headway of the connecting route. By comparing the median interchange time with the headway of the connecting routes (the legend on the right side provides the route ID of the connecting routes), the passenger experience provided by those bus-to-bus connections can be evaluated further. For example, the dots in the blue circle under the red diagonal line suggest that these bus routes provide good (or at least better than random) connecting services while the dot in the blue circle above the red diagonal line suggest that those bus routes (Routes P13 and 356) provide poorer connecting services. Thus targeted improvements

could be made to coordinate the timetable. For this example on Route 185, the connecting services are fairly good as the median interchange times are approximately half the headway of the connecting routes.

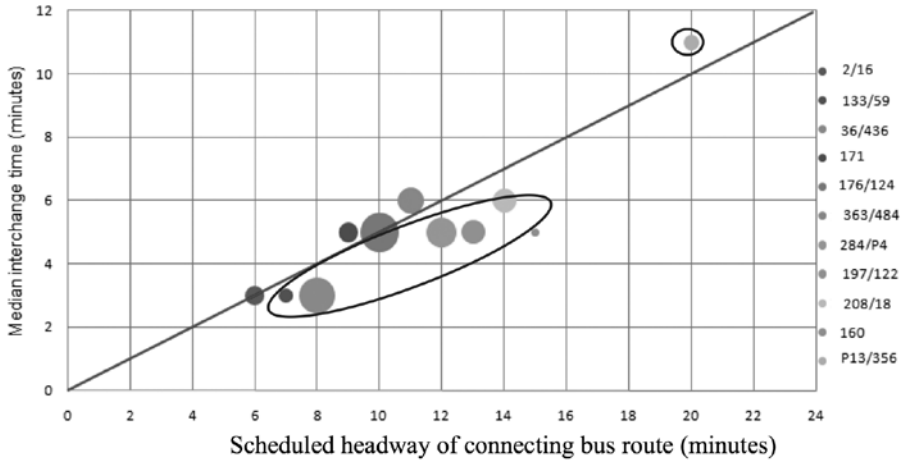


Figure 6. Relationship between connecting routes' headway and interchange time (Route 185)

Conclusions

This research has examined the feasibility of using the ADCS archived data to analyze bus passengers' travel behavior using data from TfL as an example. More reliable and comprehensive information enables public transport managers and planners to understand both their systems and customers more thoroughly, which may lead to significant changes in the effectiveness and efficiency of public transit services in the long term.

The first step in this process is to infer the origins for bus passengers by matching the smart card boarding transaction times with the AVL data. It then implements the trip-chaining methodology to infer each bus passenger's alighting location. The origin and destination inference results were then compared with the BODS manual survey data, which is the first known attempt to validate the automatic inference results against large-scale survey results. Finally, this research demonstrates potential applications of the ADCS archived data to bus network planning, with a focus on daily ridership variations and interchange time analysis, and it extends

the measurement of mobility and service performance to weekend days, for which transit planners generally have very little information (Wang 2010).

Recommendations for Future Research

Some directly-related topics for further research are recommended below:

- Disaggregate analysis of both supply and demand. On the supply side, there is the potential to "optimize" equipment use by analyzing operational data and passenger-load information. By combining these operational performance data with better demand side data, using straightforward applications such as those described in this research, it should eventually be possible to improve our understanding of the behavior of public transport users. The analysis of individual user behavior will provide additional information to transit planners on the habits of users: departure times, preferred origins and destinations, preferred routes, etc.
- Linking system usage to home addresses, access behavior also can be better understood, for instance how individuals change their behavior with weather or with the impact of improved customer information systems.
- Using cluster analysis, different user patterns can be identified and clustered into similar groups. Currently, the automatically collected data do not contain information about travel purposes, but by identifying typical temporal patterns of boardings for smart cards of similar classes, it may be possible to partition card users into commuters, students and possibly seniors who travel less than others. If the smart card number is tracked over time, the survival model of transit users and retention of different ticket types can be analyzed, which would provide longitudinal information about the network use and better information for fare planning and revenue analysis.

Acknowledgments

This research was conducted as part of a research agreement between TfL and MIT and is based on the M.S. thesis by Wei Wang under the supervision of Prof. Nigel Wilson and Mr. John Attanucci from MIT, who also provided content suggestions for and editing of the final paper. The financial support of TfL is appreciated. The authors thank Ms. Rosa McShane, Mr. John Barry, and all of TfL for their significant contribution to this research.

References

Bagchi, M., and P. R. White. 2004.The potential of public transport smart card data. *Transport Policy* 12: 464-474.

Barry, J., R. Freimer, and H. Slavin. 2008. Using entry-only automatic fare collection data to estimate linked transit trips in New York City. Transportation Research Board 2008 Annual Meeting CD-ROM, Washington, D.C.

Cui, A. 2006. Bus passenger origin-destination matrix estimation using automated data collection systems. MS Thesis, Massachusetts Institute of Technology, Cambridge.

Guo, Z., and N. H. M. Wilson. 2010. Assessing the cost of transfer inconvenience in public transport systems: A case study of the London Underground. *Transportation Research Part A*, publication pending.

Hardy, N. 2009. iBus benefits realization workstream: Method & progress to date. 16th ITS World Congress and Exhibition, Stockholm, Sweden.

Jang, W. 2010. Travel time and transfer analysis using transit smart card data. Transportation Research Board 2010 Annual Meeting CD-ROM, Washington, D.C.

Mayor of London. 2009, Transport for London – Factsheet, Available at: http://www.tfl.gov.uk/assets/downloads/corporate/transport-for-london-factsheet-july-2009.pdf.

Transport for London, Available at: http://www.tfl.gov.uk/corporate/media/newscentre/archive/15260.aspx.

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.

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): 376-387.

Wang, W. 2010. Bus passenger origin-destination estimation and travel behavior using automated data collection systems in London, UK. SM Thesis, Massachusetts Institute of Technology, Cambridge, MA.

About the Authors

Wei Wang (*winniewang@worldbank.org*) received an S.M. degree from the Department of Civil and Environmental Engineering at the Massachusetts Institute of Technology (MIT) and currently is a Junior Professional Associate in the Transport Sector of the World Bank Group. Her areas of interest include public transportation, Intelligent Transportation Systems, and transportation planning and demand analysis.

John P. Attanucci (*jattan@mit.edu*) is a lecturer and research associate in the Department of Civil and Environmental Engineering at MIT. He received a B.S. degree from the Department of Civil Engineering at Cornell University and an S.M. degree from the Department of Civil and Environmental Engineering at MIT. He specializes in public transportation management, fare policy, information technology, and transit planning and operations.

Nigel H.M.Wilson (*nhmw@mit.edu*) is Professor of Civil and Environmental Engineering at MIT. He earned S.M. and Ph.D. degrees from the Department of Civil and Environmental Engineering at MIT. His research interests include public transportation, transportation system design, and new transportation systems.