Passenger Flow Estimation and Characteristics Expansion

Prepared by:

Rabi G. Mishalani and Mark R. McCord

The Ohio State University

Prepared for:

The Ohio Department of Transportation,

Office of Statewide Planning & Research

State Job Number 134752

April 2016

Final Report

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA/OH-2016/7		

4. Title and Subtitle	5. Report Date
Passenger Flow Estimation and Characteristics Expansion	**April 2016**
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Rabi G. Mishalani **Mark R. McCord** (ORCID ID 0000-0002-6293-3143)	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
The Ohio State University **Civil, Environmental and Geodetic Engineering** **483E Hitchcock Hall** **2070 Neil Avenue** **Columbus, OH 43210**	
	11. Contract or Grant No.
	SJN 134752

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Ohio Department of Transportation **1980 West Broad Street** **Columbus, Ohio 43223**	**Final Report**
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

The objectives of this study are to investigate the estimation of bus passenger boarding-to-alighting (B2A) flows using Automatic Passenger Count (APC) and Automatic Fare Collection (AFC) (fare-box) data for general applications and for the specific purpose of expanding socioeconomic and travel (SE&T) characteristics collected though onboard interviews or questionnaires.
APC and AFC data are already being collected by transit agencies for other purposes. Therefore, if these data sources can be exploited to provide sufficiently good estimates of B2A flows, the practice of collecting B2A flows with costly and labor-intensive onboard B2A surveys could be reduced or eliminated, at least for some applications.

17. Keywords	18. Distribution Statement
Boarding-to-alighting, B2A, passenger count, fare-box	**No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161**

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	**Unclassified**	**52**	

Form DOT F 1700.7 (8-72) **Reproduction of completed pages authorized**

Passenger Flow Estimation and Characteristics Expansion

Prepared by:

Rabi G. Mishalani and Mark R. McCord,

The Ohio State University

April 2016

Prepared in cooperation with the Ohio Department of Transportation

and the U.S. Department of Transportation, Federal Highway Administration

Contents

1. Objective, Conclusions, and Report Organization

The objectives of this study are to investigate the estimation of bus passenger boarding-to-alighting (B2A) flows using Automatic Passenger Count (APC) and Automatic Fare Collection (AFC) (fare-box) data for general applications and for the specific purpose of expanding socioeconomic and travel (SE&T) characteristics collected though onboard interviews or questionnaires. APC and AFC data are already being collected by transit agencies for other purposes. Therefore, if these data sources can be exploited to provide sufficiently good estimates of B2A flows, the practice of collecting B2A flows with costly and labor-intensive onboard B2A surveys could be reduced or eliminated, at least for some applications.

Extensive empirical investigation lead to three main conclusions:

- Expensive and labor-intensive B2A survey data collected at typical sample sizes employed in practice offer little, if any, improvement in the accuracy of either disaggregate bus trip-level B2A flow matrices or more aggregate time-of-day period-level matrices. At both the bus trip- and period-level, estimates obtained from readily available APC and AFC data, without introducing B2A survey data, are seen to be better than estimates obtained when using B2A survey data alone, and very nearly as good as estimates obtained when combining B2A survey data with APC data.
- The "errors" in the estimated B2A flow matrices that are used to expand the SE&T data are attenuated in the resulting "errors" in the expanded SE&T characteristic proportions.
- B2A survey data collected at sample sizes typical of those seen in practice offer little, if any, systematic improvement for expanding SE&T characteristic data, compared to using B2A flow matrices estimated from APC and AFC data that are already being collected on an ongoing basis.

The study involves using empirical B2A flow and SE&T data manually collected on Central Ohio Transit Authority (COTA) routes and automatically collected APC and AFC data on the same routes. B2A flows are determined or estimated in multiple ways, and the resulting flows are compared to two different reference B2A flows considered to represent the ground-truth most accurately. In addition, expanded passenger SE&T characteristic proportions based on different B2A flow estimates are compared to two different reference characteristic proportions, again considered to best incorporate sampling bias corrections. It was originally envisioned that the B2A flow estimates would be compared to "ground-truth" B2A flow data collected with onboard state-of-the-practice procedures at sample rates greater than those commonly used in practice. The large sample rates were implemented as a complement to the rates considered in the AECOM memorandum to COTA (AECOM, 2014) to also allow an investigation of the sensitivity of the results to the sample rate. Similarly, it was originally envisioned that the SE&T characteristics would be compared to characteristics expanded from these "ground-truth" B2A flow data using extensions and improvements to the state-of-the-art procedure used in the AECOM memorandum. However, data quality analysis revealed that the manually collected B2A flow data could not be considered valid depictions of ground-truth B2A flows. As a result, the various B2A flow estimates are compared to two reference B2A flows, and expanded SE&T characteristics are compared to two reference characteristics, as described below. Nevertheless, the large B2A survey data sampling rates employed in this study did reveal the range over which B2A survey data does not particularly contribute to marked improvements over the use of readily available APC and AFC data for both estimating B2A flows and expanding SE&T characteristics based on B2A flows.

In light of the above objectives, the remainder of this report is organized into five additional sections and six appendices. Section 2 describes the data collected, and Section 3 presents the various B2A flow matrices used in this study. Section 4 describes the comparisons of the various B2A flows as determined or estimated with various sources of data and approaches and interprets the results. Comparisons are made at both the bus trip- and period-levels. Section 5 describes the comparisons of the various SE&T characteristics expanded using various period-level B2A flows and draws conclusions from these comparisons. The overall conclusions and directions for future research are summarized in Section 6. In addition, certain details pertaining to different aspects of the study are included in Appendices A through F.

2. Data

Four types of data on several COTA routes were collected and analyzed as part of this project. The various datasets required multiple types of processing to address measurement errors and other types of problems identified through extensive data quality investigations. In what follows the various datasets, the issues observed, and the processing procedures developed to render them useful for subsequent analyses in this study are described.

Onboard socioeconomic and travel (SE&T) characteristics survey data (OBS): Passenger demographic, socioeconomic, travel, and location related data were collected during April, May, September, and October 2013. These data were collected by ETC Institute, a leading data collection company subcontracted by OSU to administer this effort using state-of-the-art onboard, tablet-based personal interview survey techniques. In administering the data collection, ETC Institute applied quality control measures whereby the reliability of the data collectors was monitored and corrective actions were taken when necessary. In addition, ETC applied post-data collection quality control tests whereby data not meeting certain criteria were flagged and eliminated. After the completion of quality control data clean-up, the OBS captured data for 7,987 passengers and 12,512 passenger trips (many interviewed passengers conducted round-trips).

Boarding-to-alighting passenger flow survey data (B2A-S): Stop-to-stop B2A passenger flow data collection commenced prior to the OBS data collection and continued to overlap in time with the OBS data collection. Simultaneous data collection on the same route and buses was avoided. The B2A-S data were also collected by ETC Institute using a separate onboard survey technique, where data collectors distributed cards to boarding passengers and collected the cards from alighting passengers. Card identification codes were scanned and automatically geocoded by boarding location and alighting location for the purpose of inferring boarding and alighting stops. As in the case of the OBS data, ETC Institute applied field quality control measures and corrective actions. In addition, ETC applied post-data collection quality control tests whereby data not meeting certain criteria were flagged and eliminated. The OSU team identified additional data quality issues, mostly related to passengers' direction of travel, and applied necessary corrections. The OSU team also mapped the surveyed passengers to bus trips by route-directions based on the block ID information and passenger boarding times. The B2A-S data captured the boarding-to-alighting stop pairs of 45,324 passengers traveling on 2,420 bus trips.

To assess the quality of the B2A-S data on bus trips where both B2A-S and Automatic Passenger Count (APC) data (see below) are available, the boarding and alighting counts reflected in the B2A-S data are compared to the APC data by stop. Based on these comparisons, it was determined that the

representativeness of the B2A-S data is poor at the bus trip and stop levels. This determination was made after observing generally poor correspondence between B2A-S and APC boarding and alighting activity at stops and comparisons of quantified correspondences to quantified correspondences produced on OSU's Campus Transit Lab (CTL). (OSU's CTL (CTL, 2015) is a living lab well suited for high-fidelity data collection, including onboard B2A surveys.) B2A-S data were subsequently aggregated into stop segments and time-of-day periods in an attempt to increase the representativeness of the passenger flows. Furthermore, criteria to screen B2A-S data for large errors were developed. These criteria and their implementation are discussed in more detail in Section 4 and Appendix A. In addition, the OSU team mapped passengers to the four time-of-day periods defined by COTA based on passengers' boarding times. Finally, because not all data collectors started or finished their data collection at the terminals of routes, the passenger flows on a large number of bus trips are incomplete. B2A-S data exhibiting this characteristic are referred to as partial trip data. Appendix B presents the approach adopted to correct B2A-S partial trip data.

Automatic Passenger Count data (APC): The APC data are collected by COTA using an APC technology installed on approximately 20% of the bus fleet. APC data spanning two trimesters – January to April and May to August 2013 – and one month – September 2013 – were provided by COTA. These data encompass 97,800 bus trips, on which a total of 2,175,125 passengers were observed (based on the average of the boarding and alighting counts on each bus trip). The OSU team investigated the quality of the APC data by comparing the total boarding and total alighting counts on each bus trip. APC data on bus trips exhibiting large discrepancies between the trip-level boarding and alighting counts reflect appreciable measurement errors and, as a result, are eliminated from further consideration. Depending on the trimester (two) or month (one), 35% to 41% of bus trips and 27% to 30% of passengers are eliminated based on this screening analysis.

Automatic Fare Collection data (AFC): The AFC (fare-box) transaction data are collected by COTA when electronic fare media are used to pay the fare. The technology used relies on a separate Automatic Vehicle Location (AVL) system than the one used to identify the locations and stops of the APC data. AFC transaction data from January through May 2013 and in September 2013 were provided by COTA and included a total of 8,208,785 passenger transaction records. These data are collected when boarding passengers swipe on using any fare payment medium. As a result, the boarding stop is included in the transaction data. However, since alighting passengers do not swipe off on COTA, the destination stop is not included in the AFC transaction data. As part of this project, destination stop locations were inferred from additional transaction activities using the same fare medium (more detail is provided subsequently in Section 3).

Prior to applying the alighting location inference, it was determined that the locations of the AFC transactions were inconsistent and unreliable. However, the timestamps seemed to be correct. This situation was confirmed to be the case by COTA. Since the locations in the AFC and APC data are collected through two separate and independent AVL systems, the two datasets were matched based on timestamps for each route-direction. The accurate stop locations of the APC data were then migrated to the corresponding AFC data. Because only approximately 20% of the COTA bus fleet is equipped with APC technology, only 18% to 20%, depending on the month, of the AFC transactions could be corrected in this manner. As discussed in more detail in Section 3, the inference of alighting stops relies on transfer and return-tip AFC transactions. As a result, once the inference is applied only less than 4% of the passengers with AFC boarding stop information have an associated alighting stop information.

The four datasets are used in this study to compare B2A flows and to assess the quality of SE&T characteristic expansions based on various B2A flows. The B2A flows are compared at the bus trip and time-of-day period levels in Sections 4.1 and 4.2, respectively. For the purpose of comparing the B2A flows at the trip level, bus trips are identified for which B2A-S, APC, and AFC data are all available for full (complete) bus trips after screening the B2A-S and APC data for errors and applying the AFC boarding stop location correction and alighting stop inference. Only 34 bus trips across all routes, directions, and times of day are thus identified.

Screening for B2A-S errors could not be applied at the period level, since the quantity of remaining B2A-S data would not be sufficient to conduct any meaningful comparisons. (However, the large amount of APC data allowed screening for poor quality APC data at the period level.) To ensure that a sufficient amount of B2A-S data are available to investigate the effect of B2A-S bus trip sample size on the results, both full and corrected partial bus trips are used, and only routes and time-of-day periods with sufficient B2A-S data are considered. (Boarding timestamps are used to aggregate trip-level B2A flows to a time-of-day period.) The time-of-day periods with sufficient B2A data are 9 am to 3 pm and 3 pm to 7 pm, periods 1 and 2 as defined by COTA. It is noted that time-of-day periods are generally determined based on volumes. Since this study is focused on B2A flow patterns, it would be interesting to determine periods based on such patterns and to refine the time-of-day periods for subsequent investigation in future studies. (A methodology for identifying and determining homogenous time-of-day periods based on B2A flow patterns is presented in Ji et al. (2011).) Similarly, unscreened B2A-S data and period-level B2A flows are considered in the investigation of the expansion of SE&T characteristics presented in Section 5.

Table 1 presents the sample sizes of all four datasets for the routes considered in the period-level B2A comparisons and in the SE&T characteristics expansions and their comparisons. (Of the 34 trips considered in the trip-level comparisons of Section 5, 31 come from the same set of routes. The other three trips are on routes 81 and 84.) The questionable validity of the unscreened B2A-S data in representing the ground-truth B2A flows raises challenges regarding the assessment of the accuracies of the B2A flow matrices and SE&T characteristics expansions. These challenges are addressed in Sections 4 and 5.

Table 1: Sample sizes of datasets for the routes considered in the period-level B2A flow and SE&T characteristics comparisons

		Route							
Data		1	2	3	7	9	10	11	Total
OBS	Passenger Trips	972	1262	385	207	193	959	235	4,213
	Passengers	595	821	246	129	120	648	148	2,707
APC	Bus Trips	4067	6021	1017	1,799	1,088	3154	1,158	18,304
	Passengers	167,947	271,338	21,017	42,383	17,892	126,387	24,759	671,723
B2A-S	Bus Trips	229	263	79	66	71	255	94	1,057
	Passengers	5,457	6,705	1,259	1,512	966	5,447	1,498	22,844
	Full Trips	90	87	51	50	25	108	52	463
	Full Trip Passengers	3,146	3,444	1,016	1,309	447	3,359	1,134	13,855
	Partial Trips	139	176	28	16	46	147	42	594
	Partial Trip Passengers	2,311	3,261	243	203	519	2,088	364	8,989
	Corrected Partial Trip Passengers	5207	7433	518	410	745	4,546	897	19,754
AFC[1]	Bus Trips	4,376	7,249	1,835	2,610	2,108	2,034	1,887	22,099
	Passengers	71,285	115,521	11,332	19,573	11,613	58,961	14,528	302,813

[1]The reported values correspond to results following the application of the alighting stop inference procedure.

3. B2A Flow Matrices

Boarding-to-alighting flow estimates are organized into B2A flow matrices either for individual bus trips or for bus route, direction, and time-of-day period combinations. (The route-direction-period matrices will often be referred to as "period-level" matrices for simplicity.) Trip-level matrices depict the B2A flows of passengers travelling on individual bus trips. Period-level flows depict the B2A flows of passengers traveling on any trip on the route-direction with boarding times falling within the time-of-day period. Since the period-level matrices are developed from the trip-level matrices, trip-level matrices are the more elemental of the two for investigating the accuracy of B2A flow estimation. On the other hand, period-level B2A flow matrices would typically be of more interest for practical applications – e.g., the expansion of SE&T characteristics considered in this study. Therefore, both trip- and period-level estimations are investigated.

The passenger flow matrices are developed to reflect passenger flow probabilities, or proportions, between pairs of boarding and alighting stop segments. Passenger flow probabilities are determined by

dividing passenger flow volumes for boarding-to-alighting pairs by the total passenger flow volume in the matrix. In this way, the matrix represents the probability that a random passenger traveling on the bus trip or on the bus route-direction during the time-of-day period would travel from the boarding stop to the alighting stop indicated by the boarding-to-alighting cell. Considering flow probabilities places the focus on the spatial dimension of flow in the matrices, rather than on flow volumes. Given that the boarding-to-alighting survey data and the AFC data are samples of the passenger volumes, the interest is in whether these samples can represent the spatial flow patterns. Moreover, as will be seen below, passenger flow probabilities are the fundamental elements used to expand SE&T characteristics.

The stop-to-stop passenger flow probabilities are aggregated to reflect passenger flows between pairs of stop segments. Stop segments group consecutive stops in a larger set of stops intended to represent a geographical region. The stop segments used in this study are those used by AECOM in a study for COTA (AECOM, 2014). (Other approaches for grouping stops into segments – e.g., McCord et al. (2012) – could be explored as part of future search.) Grouping into stop segments diminishes the effect of slight data collection or estimation errors in the B2A flow estimates (e.g., recording or estimating the boarding or alighting to have occurred a stop or two upstream or downstream).

The various B2A flow matrix estimates developed using different data sources and estimation approaches are defined and summarized in Table 2 and described in what follows. The *B2A-S* matrix for a bus trip is formed directly from the manually collected boarding-to-alighting survey data for the bus trip. The stop-to-stop passenger flow volumes on a bus trip are normalized by the total passenger volume on the trip to produce stop-to-stop flow probabilities. The stop-to-stop flow probabilities are then aggregated into segment-to-segment flow probabilities by summing the stop-to-stop passenger flow probabilities for all stop-to-stop pairs encompassed in the segment-to-segment pair. The *B2A-S* matrix for a route, direction, and period is formed by summing all the bus trip level volume matrices in the route, direction, and time-of-day period considered, normalizing to produce flow probabilities, and aggregating the stop-to-stop probabilities into segment-to-segment probabilities. (Summing bus trip level volume matrices and then normalizing, rather than normalizing bus trip level matrices and summing, allows bus trips with larger volumes to have a greater impact on the period level matrix.)

The *X%B2A-S* matrix is formed by randomly generating a sample of passengers from the boarding-to-alighting survey data that is approximately X% of the target daily passenger ridership for the route, direction, and period. (Daily passenger ridership is considered in practice to represent "100%" of the population. Larger samples could be obtained by sampling trips multiple times on different days. Accounting for this temporal variability can lead to a more accurate B2A flow characterization, as has been shown to be the case in an investigation using data from this study.) The *X%B2A-S* matrix is used to investigate the benefits of the improved representativeness of the matrix associated with larger sample size, since the larger sample would require larger data collection costs. A random sample of passengers could be obtained by independently generating passenger boarding-to-alighting flows from the *B2A-S* matrix or by generating sets of passenger boarding-to-alighting flows on bus trips. The latter is used in this study, since it is more consistent with present data collection practice. In an investigation using data from this study, sampling by passengers, rather than sampling by bus trips has been found to result in more accurate B2A flow matrices for the same total number of passengers in the sample.

Table 2: Descriptions of B2A flow matrices

Notation	Data	Method	Comments	Purpose for this Study
B2A-S[1,2]	B2A survey (all data)	Direct observation	Requires special costly and labor intensive data surveys; subject to B2A survey measurement errors	Intended to represent trip- and period-level ground-truth B2A flow matrices in the absence of other data sources and used to determine the most accurate expanded characteristics in the absence of other data sources
X%B2A-S[2]	Sampled B2A survey	Direct observation; B2A surveyed trips are sampled to achieve X% of average ridership for time-of-day period (X% = 10%, 25%, 50%, 75%, 100% in this study)	Requires special surveys; subject to B2A survey measurement errors and sampling errors	Represents a sample of the period-level ground-truth B2A flow matrix in the absence of other data sources, and used to determine expanded characteristics in the absence of other data sources
AFC[1,2]	AFC	Direct observation of boarding stop from fare medium transaction information and inferred alighting stop from fare medium transaction information regarding transfers or return trips	Uses information collected by transit agencies for fare collection purposes; gaining popularity in determining B2A flows where fare medium data are available; alighting stop inference subject to travel assumptions; estimates also subject to stop location errors in this study; matched to APC data in this study to correct for location errors leading to a very small sample of usable AFC data	Represents trip- and period-level B2A flows when only AFC data are available, and used to determine expanded characteristics when only AFC data are available
Null[1,2]	None	Equal bus stop-level probabilities are determined as the reciprocal of the number of feasible stop-to-stop B2A pairs. Stop-level probabilities are aggregated into segment-level probabilities by aggregating stop-to-stop pairs into segment-to-segment pairs.	Represents equal flow probabilities between all feasible stop-to-stop pairs. Used as a benchmark for B2A flows without any prior knowledge of flow patterns.	Represents flow matrix where a randomly selected passenger is equally likely to travel between any feasible B2A stop-to-stop pair and used (a) as a benchmark B2A estimate (b) as seed (or base) in estimating certain B2A flows, and (c) to address the non-structural zeros of other matrices used as seeds
IPF(Null)[1,2]	APC	Iterative Proportional Fitting (IPF) method using the uninformative Null matrix, where a random passenger is assumed to be equally likely to travel on any feasible OD pair, as a seed matrix	State-of-the practice method using boarding and alighting count data when no prior or model generated informative matrix is available to serve as a seed; subject to methodological inaccuracies and count measurement errors	Represents trip- and period-level B2A flows when only APC data are available, and used to determine expanded characteristics when only APC data are available

[1]Used at trip-level.

[2]Used at period-level.

Table 2, continued: Descriptions of B2A flow matrices

Notation	Data	Method	Comments	Purpose for this Study
IPF(B2A-S)[1,2]	APC, B2A survey (all data)	IPF method using the period-level B2A flow matrix as a seed; seed matrix is an aggregation of all trip-level B2A flow survey data for time-of-day period	Uses boarding and alighting count data when a period-level B2A flow matrix determined from B2A survey data is available to serve as a seed; subject to methodological inaccuracies, count measurement errors, and B2A survey errors	Believed to represent the most accurate B2A flow matrix estimates for this study; considered a reference, ground-truth matrix for trip- and period-level B2A flow estimates comparisons, and used to determine expanded characteristics believed to represent the most accurate characteristics
IPF(X%B2A-S)[1,2]	APC, sampled B2A survey	IPF method using a sampled period-level B2A flow matrix as a seed; seed matrix is the X% B2A-S matrix described above	Uses boarding and alighting count data when a period-level B2A flow matrix determined from B2A survey data is available to serve as a seed; B2A survey data reflect X% of average time-of-day period ridership; subject to methodological inaccuracies, count measurement errors, B2A survey measurement errors, and B2A sampling errors	Represents trip- and period-level B2A flows and used to determine expanded characteristics when APC data and B2A flow survey data sampled to achieve X% of average period ridership are available
IPF(AFC)[1,2]	APC, AFC	IPF method using the period-level AFC B2A flow matrix as a seed; seed matrix is the AFC matrix described above	Uses boarding and alighting count data when a period-level B2A flow matrix determined from AFC data is available to serve as a seed; subject to methodological inaccuracies, count measurement errors, AFC inference errors, and low AFC sample size in this study	Represents trip- and period-level B2A flows and used to determine expanded characteristics when APC and AFC data are available

[1]Used at trip-level.

[2]Used at period-level.

For a given route, direction, and time-of-day period, bus trips are randomly generated one at a time from the set of bus trips where boarding-to-alighting survey data were collected until the observed number of passengers surveyed first surpasses X% of the target daily passenger ridership for the route, direction, and period. The boarding-to-alighting flow volumes are aggregated by segment-to-segment pair and across all trips on the route during a time-of-day period to form a matrix for the route, direction, and period and then normalized to produce the flow probability matrix. The random generation is repeated ten times using a different random seed each time to form ten different sets of *X%B2A-S* matrices for each route, direction, and period.

The *AFC* matrix is formed by inferring alighting stops for boarding transactions recorded by automatic fare payment media. Specifically, the stop location where a transfer takes place is used to infer the alighting stop on a bus trip leading to that transfer, and the boarding stop location of a return passenger trip is used to infer the alighting stop on a bus trip that where no subsequent transfer transaction is available. Once the individual boarding-to-alighting movements are determined, bus trip- and period-

level matrices are determined by route, direction, and period using the same procedures that were used when determining *B2A-S* matrices.

The Null matrix is formed by considering equal bus stop-to-bus stop flow probabilities whose values are set to the reciprocal of the number of feasible B2A pairs on the bus route and direction during the time-of-day period. (A feasible cell is one where passenger flow could be observed in the boarding-to-alighting pair represented by this cell.) In this way, the Null matrix represents the case where a random passenger is equally likely to travel along any feasible boarding stop-to-alighting stop pair. The Null matrix is, therefore, considered "uninformative". The stop-to-stop Null matrix is aggregated into a segment-to-segment Null matrix by adding the passenger flows in the matrix for all stop-to-stop pairs contained in the segment-to-segment pair. As a result, the flow probabilities are generally unequal in the Null segment-to segment bus trip matrix. Since the Null stop-to-stop matrix is the same for each route-direction bus trip in a period, the route-direction-period Null matrix is the same as the bus trip-level matrix for a route-direction-and period. (Since bus trips in different periods may serve different numbers of stops, the Null matrix can differ by period.)

The *IPF(Null)*, *IPF(B2A-S)*, *IPF(X%B2A-S)*, and *IPF(AFC*) matrices are B2A matrices estimated using the state-of-the-practice Iterative Proportion Fitting (IPF) procedure (Ben-Akiva et al., 1985; Simon and Furth, 1985; McCord et al., 2010; Mishalani et al., 2011), taking as input trip-level APC data and the corresponding period-level *Null*, *B2A-S*, *X%B2A-S*, and *AFC* matrices, respectively, as seed matrices. The number of passengers boarding at a specified stop must equal the sum, over all alighting stops, of the B2A flows originating at the boarding stop. Similarly, the number of passengers alighting at a specified stop must equal the sum, over all boarding stops, of the B2A flows destined to the alighting stop. In the IPF method, a seed matrix is iteratively adjusted until the matrix converges to one whose elements satisfy these relations for all boarding and alighting stops, where the APC data are used to provide the boarding and alighting volumes.

For the IPF method to converge, it is necessary that the sum of all the boarding volumes is equal to the sum of all the alighting volumes. (This implies that the number of passengers who board on a bus trip is equal to the number of passengers who alight on the bus trip.) Because of APC errors and, to a lesser extent, of some passengers alighting on a bus trip in one direction, staying on the bus through the terminal, and alighting on a bus trip in the other direction, the sum of the APC-based bus trip boardings will not always equal the sum of the APC-based bus trip alightings. Therefore, a procedure is first applied to "balance" the boarding and alighting data (Furth et al., 2005). (In addition to ensuring that that total boarding and alighting volumes are equal, the procedure ensures that the balanced boarding and alighting volume vectors imply no negative passenger loads between any pair of stops.)

The seed B2A matrix, which is used as input along with the APC-based boarding and alighting counts in the IPF method, can be considered an initial estimate of the B2A seed matrix. The same seed period-level matrix is used for all bus trips on the route, in the direction and during the time-of-day period considered. *Null*, *B2A-S*, *X%B2A-S*, and *AFC* matrices are used as seeds to investigate the quality of using APC data in combination with various other data sources.

The relatively small sample sizes in the B2A survey and the inability to infer alighting stops for many of the observed AFC boardings (because of the data difficulties discussed above) can lead to some segment-to-segment flow probabilities equal to zero in the period-level *B2A-S, X%B2A-S,* and *AFC* matrices. Having "nonstructural zeroes" in these matrices is troublesome when the matrices are used to

expand the SE&T characteristics. When used in the expansion procedure, zero values in the B2A matrix will lead to zero values in the expanded characteristics for the corresponding segment-to-segment pair, even though the OBS data may exhibit non-zero values for this pair. In addition, when used as seed (base) matrix for the IPF procedure, zero values will lead to values of zero in the IPF output. These matrices would then suffer from the same difficulty when used to expand SE&T characteristics. Therefore, a procedure, described in Appendix C, was developed to replace these nonstructural zeros with reasonable nonzero values.

4. B2A Flow Comparisons

As discussed above, period level B2A matrices are constructed from the bus trip-level B2A matrices. Therefore, investigating the accuracy of the various methods and datasets at the more elemental bus trip level would provide a better indication of B2A estimations accuracy. On the other hand, the period-level matrices are typically more useful for practical applications. Therefore, accuracy in estimating both trip-level and period-level matrices is investigated. Because of the errors in the manually collected B2A survey data, a subset of the bus trips that appear to have better quality B2A-S data are used in the trip level analysis. In Appendix A, the screening procedures developed to determine a set of bus trips with what is believed to be relatively good quality B2A survey data are described. After applying these screening procedures and matching the remaining trips to those for which AFC alighting inference could be made, there were too few trips (34) remaining to develop period-level matrices. Therefore, as discussed previously, all the bus trips that pass the APC screening criteria, reflect corrected boarding stop locations in the AFC transactions with inferred alighting stops, and correspond to the relevant periods are used in the period-level analyses.

As mentioned above, the original intent was to compare estimated B2A flow matrices to ground-truth matrices determined from large quantities of B2A survey data. However, the difficulty in collecting high-fidelity ground-truth B2A flow data for such a large-scale empirical study required a change in perspective to one of using reference, rather than ground-truth, B2A flow matrices as the basis for comparing B2A flow estimates. The *B2A-S* and *IPF(B2A-S)* flow matrices are used as references for the comparisons as possible representations of the most accurate depictions of ground-truth B2A flows. The *B2A-S* matrices represent flows obtained from the manually administered onboard B2A survey. The screening of bus trips developed and implemented increases the confidence that the *B2A-S* matrices are more valid representations of flow patterns on the remaining bus trips than on the entire set of bus trips. However, even after screening the *B2A-S* matrices, the multiple errors seen make the validity of these matrices questionable. Moreover, an insufficient number of bus trips remain after the screening to be meaningful to develop period-level matrices as estimates or as seed (base) matrices for use with APC data. The *IPF(B2A-S)* matrices refine the *B2A-S* matrices by using the APC boarding and alighting data. However, the refinements are limited by inaccuracies in the IPF method and by measurement errors in the APC data. Therefore, comparisons are conducted considering both references.

Matrices estimated with different datasets and approaches are compared to the reference matrices using the Hellinger Distance (*HD*) measure. This commonly used measure maps the multi-dimensional distance between a pair of matrices into a scalar number such that two matrices with a lower *HD* value are considered more similar to each other than are two matrices with a higher *HD* value. Specifically, the *HD* measure is given by:

$$HD = \sqrt{\sum_i \sum_j \left(\sqrt{\hat{p}(i,j)} - \sqrt{p(i,j)} \right)^2} \quad (1)$$

where, $\hat{p}(i,j)$ represents the probability B2A flow from segment *i* to segment *j* determined or estimated with different datasets and approaches, and $p(i,j)$ represents the probability B2A flow from segment *i* to segment *j* of one of the two reference B2A flows.

4.1 Trip-level B2A Flow Comparisons

A bus trip is considered in the empirical investigation of B2A estimation at the bus trip-level if all of the following datasets are available for the trip:

- Flows from the B2A survey that pass the screening tests described in Appendix A to increase the confidence in the validity of the B2A survey data for the trip.
- APC data that pass the screening tests developed in Section 2 to increase the confidence in the validity of the APC data for the trip.
- AFC data where the locations of the boarding stops have been corrected as discussed in Section 2 and the alighting stops have been inferred from subsequent AFC transaction data as discussed in Section 3.

Only 34 bus trips – across all routes, directions, and periods – have all three of these data sources. For each of the 34 bus trips, the *B2A-S*, *IPF(B2A-S)*, *IPF(AFC)*, *IPF(Null)*, *AFC*, and *Null* matrices are determined. In addition, ten *IPF(100%B2A-S)* matrices, ten *IPF(75%B2A-S)* matrices, ten *IPF(50%B2A-S)* matrices, ten *IPF(25%B2A-S)* matrices, and ten *IPF(10% B2A-S)* matrices are determined. The ten different *IPF(X%B2A-S)* matrices are formed by considering ten different random samples of *all* the B2A survey data (i.e., not just the screened data) for the corresponding route, direction, and period to form the seed (base) matrix to be used with the trip's APC data when applying the IPF procedure. Since only 34 trips – spread across routes, directions, and periods – remain after screening for data quality, samples could not be generated for a specific route, direction, and period from the screened B2A survey data. Therefore, all the B2A-S data are considered when determining the seeds (bases) for the IPF method.

The *HD* value is determined for each trip-level estimate relative to each of the *B2A-S* and *IPF(B2A-S)* matrices used as the reference. Summary statistics are presented In Tables 3a and 3b. Table 3a presents statistics when using the *B2A-S* matrix as the reference, and Table 3b presents statistics when using the *IPF(B2A-S)* matrix as the reference.

To allow a more visual comparison of the results, the mean *HD* values for the matrices obtained using the IPF method with seed matrices formed from B2A survey data are plotted in Figure 1a and 1b as a function of the sample size taken to determine the B2A seed matrix. As noted above, the sample size is given as a percentage of the average daily ridership for the corresponding route, direction, and period. Figures 1a and 1b correspond, respectively, to results obtained when using the *B2A-S* and *IPF(B2A-S)* matrices as references. It is noted that a B2A sample size of 0% corresponds to the *IPF(Null)* matrix. It is also noted that, since the *AFC*, *IPF(AFC)* and *Null* matrices do not use B2A survey data when determining or estimating B2A matrices, the *HD* values obtained for these matrices do not vary by sample size of the B2A survey data, and they are plotted as horizontal lines in the figures. This is also the case when

plotting the values corresponding to either the *B2A-S* and *IPF(B2A-S)* matrices when either of the two is used as the reference.

Table 3a: Summary statistics of HD measure comparing bus trip B2A flow matrices determined or estimated from different datasets to reference flow matrix B2A-S after screening B2A-S data

	Mean	Standard Deviation	Coeff. of Variation	10th Percentile	50th Percentile	90th Percentile
IPF(All B2A-S)	0.397	0.200	0.505	0.188	0.331	0.679
IPF(100%B2A-S)	0.402	0.199	0.495	0.190	0.332	0.714
IPF(75%B2A-S)	0.402	0.199	0.494	0.189	0.333	0.702
IPF(25%B2A-S)	0.403	0.197	0.490	0.192	0.339	0.707
IPF(50%B2A-S)	0.405	0.200	0.493	0.189	0.336	0.725
IPF(10%B2A-S)	0.410	0.199	0.486	0.196	0.345	0.727
IPF(AFC)	0.413	0.206	0.499	0.196	0.352	0.717
IPF(Null)	0.424	0.207	0.488	0.197	0.374	0.726
AFC	0.616	0.212	0.345	0.329	0.625	0.835
Null	0.762	0.162	0.213	0.527	0.796	0.933

Table 3b: Summary of statistics of HD measure comparing bus trip B2A flow matrices determined or estimated from different datasets to reference flow matrix IPF(B2A-S) after screening B2A-S data

	Mean	Standard Deviation	Coeff. of Variation	10th Percentile	50th Percentile	90th Percentile
IPF(100%B2A-S)	0.033	0.029	0.865	0.002	0.025	0.068
IPF(75%B2A-S)	0.040	0.031	0.786	0.006	0.034	0.078
IPF(50%B2A-S)	0.051	0.038	0.749	0.009	0.043	0.099
IPF(25%B2A-S)	0.064	0.046	0.726	0.010	0.056	0.126
IPF(10%B2A-S)	0.088	0.059	0.674	0.014	0.082	0.168
IPF(AFC)	0.092	0.065	0.705	0.014	0.082	0.172
IPF(Null)	0.094	0.051	0.541	0.019	0.097	0.156
B2A-S	0.397	0.200	0.505	0.188	0.331	0.679
AFC	0.556	0.217	0.389	0.283	0.546	0.830
Null	0.640	0.154	0.241	0.437	0.629	0.855

The tables and figures show that, regardless of the reference considered, the *IPF(X%B2A)* and *IPF(AFC)* matrices improve the quality of the B2A flows substantially (i.e., reduce the *HD* values) compared to the uninformative *Null* B2A matrices or the B2A matrices estimated solely from AFC data. That is, using APC data with the IPF method is beneficial. Not surprisingly, the *IPF(X%B2A)* and *IPF(AFC)* matrices are closer to the reference *IPF(B2A-S)* matrices (mean *HD* values ranging from 0.033 to 0.094), which also use APC data, than to the reference *B2A-S* matrices (mean *HD* values of 0.397), which do not use APC data.

As expected, the *IPF(X%B2A)* estimates improve (i.e., exhibit lower *HD* values) with increases in sample size of the B2A survey data used to form the seed matrix. However, regardless of the reference, the difference in mean *HD* between the 0% sample size (i.e., *IPF(Null)*) and the extremely large 100% sample size is very small compared to the difference in mean *HD* between matrices estimated without using APC data (namely, *Null* and *AFC* matrices) and those estimated with APC data.

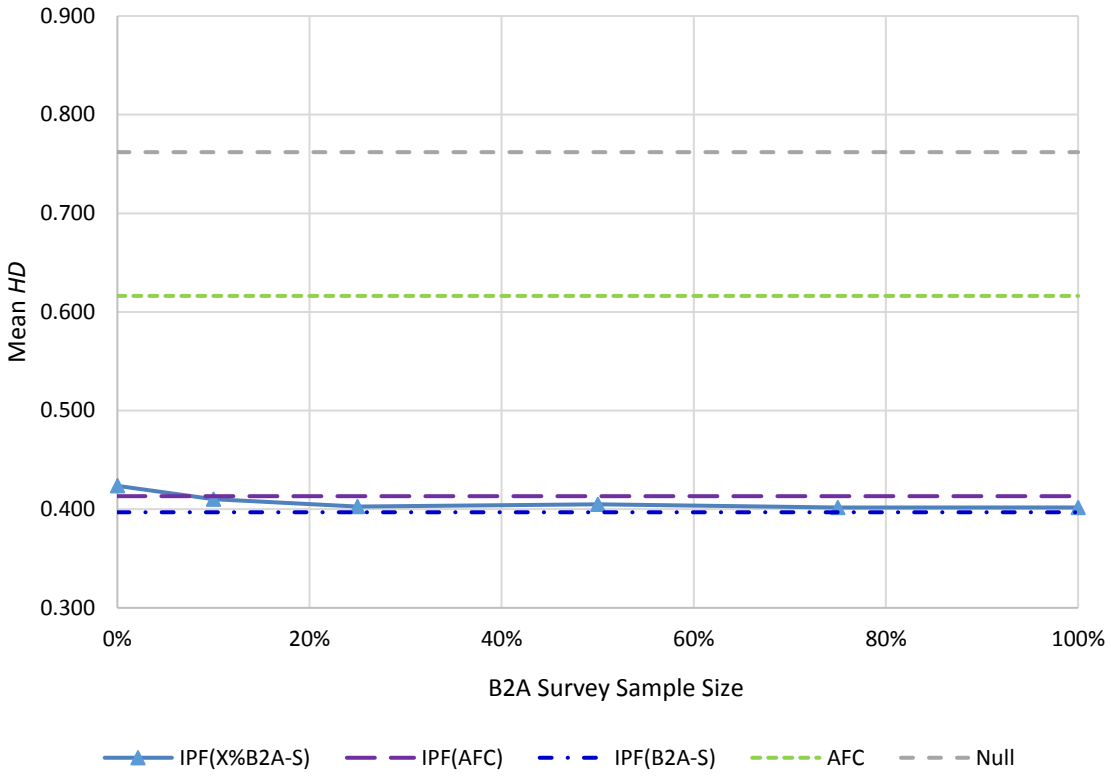

Figure 1a: Mean HD values comparing bus trip B2A matrices determined or estimated from different datasets to reference matrix B2A-S after screening B2A-S data

Of particular interest are the comparisons between the *IPF(Null)* (0% sample size) or *IPF(AFC)* matrices, which do not require any B2A survey data, and the *IPF(X%B2A)* matrices with sample sizes less than 25%, which would be considered greater than sample sizes typically collected in practice. The mean *HD* values of the *IPF(25%B2A)* estimates are less than the mean *HD* values of the *IPF(Null)* and *IPF(AFC)* estimates for both references, but only by a small amount. The mean *HD* values of the *IPF(10%B2A)* show practically no improvement, regardless of the reference. The implication is that, on average, very little improvement in accuracy is gained from the use of B2A survey data with sample sizes typically used in practice.

The plots in Figures 1a and 1b portray the trends in the mean performance of the estimation method and data sets. The standard deviation, coefficient of variation (mean divided by standard deviation), and percentile values in Tables 3a and 3b indicate that there is a fairly large amount of variability about the mean values. Therefore, the percentage of times that the *IPF(Null)*, *IPF(AFC)*, and *AFC* matrices (matrices that can be determined or estimated without the need for any onboard B2A survey data collection) are closer to the reference than are the *IPF(25%B2A-S)* matrix or the *IPF(10%B2A-S)* matrix are determined. The percentages, which are determined by considering all trips and B2A survey data samples, are presented in Table 4. The 25% B2A survey sample size is used as an upper bound on what would typically be collected in practice, whereas a 10% sample size is used to represent the performance associated with decreasing the sample from what might be considered a typical value.

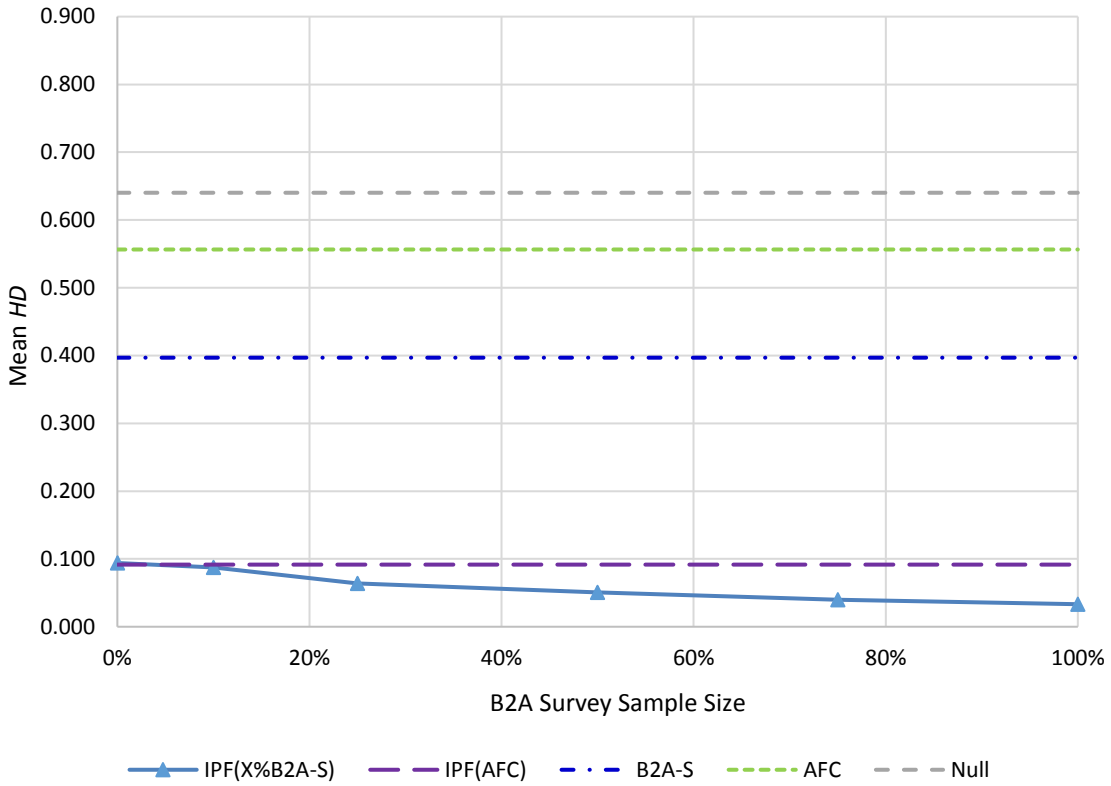

Figure 1b: Mean HD values comparing bus trip B2A matrices determined or estimated from different datasets to reference matrix IPF(B2A-S) after screening B2A-S data

Table 4: Percentage of times IPF(Null), IPF(AFC), and AFC matrices are closer than the IPF(10% B2A-S) and IPF(25% B2A-S) matrices to the reference B2A matrices

	B2A-S Reference			IPF(B2A-S) Reference		
	IPF(Null)	IPF(AFC)	AFC	IPF(Null)	IPF(AFC)	AFC
IPF(10% B2A-S)	48.0%	49.8%	23.9%	43.6%	48.6%	1.3%
IPF(25% B2A-S)	46.5%	48.2%	22.8%	31.1%	36.5%	0.5%

The approximately 50% values in most of the *IPF(Null)* and *IPF(AFC)* cells indicate that these matrices perform better than the *IPF(10%B2A-S)* and *IPF(25%B2A-S)* matrices on approximately half the bus trips. The percentages are noticeably lower (31% and 37%) when comparing to the *IPF(25%B2A-S)* matrices and considering the *IPF(B2A-S)* matrices as the reference. However, the 25% B2A sample size is considered large, and the *IPF(Null)* and *IPF(AFC)* matrices still outperform the resulting matrices over 30% of the time. When coupled with the other results in the table and the analysis of the mean performance seen in Figures 1a and 1b, the empirical analysis indicates that, at the individual bus trip level, B2A survey data with sample sizes used in practice offer little, if any, improvement over the *IPF(Null)* and *IPF(AFC)* matrices, which require no B2A survey data. On the other hand, the performance of the *AFC* matrices is appreciably worse than even the *IPF(10%B2A-S)* matrices, indicating the B2A

survey data at sample sizes used in practice produce appreciably better B2A estimates than estimates obtained when only using the AFC data, but that the real benefit is obtained by using APC data.

4.2 Period-level B2A Flow Comparisons

The 34 trips remaining after screening the trip-level data are too few to use to form period-level matrices for a route, direction, and period. Therefore, all the B2A survey data with associated block ID (used to identify bus trips), and not just the better quality, screened data, are used in the period-level analysis. The same screening criteria used in the trip-level analysis are used for the APC and AFC data. To investigate performance when using a sample of B2A survey data during a period that is equal to the average daily ridership in the period (100%B2A), it is necessary that the quantity of B2A survey data be at least this amount. Such data are available in both directions for Periods 1 (9:00 am to 3:00 pm) and 2 (3:00 pm to 7:00 pm) on Routes 1, 2, 3, 7, 9, 10, and 11 in both directions. These are, therefore, the route-direction-periods used in the period-level analysis.

For each route, direction, and period considered, the *B2A-S*, *IPF(B2A-S)*, *IPF(AFC), IPF(Null)*, *AFC*, and *Null* matrices are determined. In addition, ten different *100%B2A-S*, *75%B2A-S*, *50%B2A-S*, *25%B2A-S*, and *10%B2A-S* matrices are formed by considering ten different random samples of all the bus trips containing B2A survey data for the corresponding route, direction, and period. These matrices are used to form ten samples of the period-level B2A matrices for the route and direction and to serve as ten different seed (base) matrices to form ten different period-level *IPF(100%B2A-S)*, *IPF(75%B2A-S)*, *IPF(50%B2A-S)*, *IPF(25%B2A-S)*, and *IPF(10%B2A-S)* matrices for the route and direction.

Summary statistics of the *HD* measure obtained when comparing the various period-level matrices to the corresponding reference matrices for the route, direction, and period are presented in Tables 5a and 5b. Table 5a presents statistics when using the *B2A-S* matrix as the reference, and Table 5b presents statistics when using the *IPF(B2A-S)* matrix as the reference.

Table 5a: Summary statistics of HD measure comparing period B2A matrices determined or estimated from different datasets to reference matrix B2A-S

	Mean	Standard Deviation	Coeff. of Variation	10th Percentile	50th Percentile	90th Percentile
IPF(AllB2A-S)	0.164	0.048	0.290	0.103	0.165	0.222
IPF(100%B2A-S)	0.167	0.046	0.275	0.106	0.168	0.224
IPF(75%B2A-S)	0.169	0.046	0.270	0.109	0.170	0.226
IPF(50%B2A-S)	0.171	0.046	0.268	0.111	0.172	0.228
IPF(25%B2A-S)	0.177	0.047	0.268	0.115	0.176	0.234
IPF(10%B2A-S)	0.184	0.050	0.270	0.116	0.184	0.244
IPF(AFC)	0.189	0.048	0.254	0.130	0.192	0.247
IPF(Null)	0.202	0.053	0.264	0.132	0.199	0.271
100%B2A-S	0.084	0.057	0.675	0.029	0.069	0.151
75%B2A-S	0.104	0.059	0.564	0.044	0.091	0.173
50%B2A-S	0.128	0.059	0.464	0.063	0.117	0.203
25%B2A-S	0.171	0.072	0.421	0.091	0.154	0.265
10%B2A-S	0.230	0.086	0.375	0.125	0.221	0.349
AFC	0.308	0.106	0.343	0.217	0.277	0.492
Null	0.510	0.115	0.226	0.381	0.517	0.702

Table 5b: Summary statistics of HD measure comparing period B2A matrices determined or estimated from different datasets to reference matrix IPF(B2A-S)

	Mean	Standard Deviation	Coeff. of Variation	10th Percentile	50th Percentile	90th Percentile
IPF(100%B2A-S)	0.021	0.012	0.571	0.007	0.020	0.037
IPF(75%B2A-S)	0.026	0.014	0.546	0.009	0.025	0.046
IPF(50%B2A-S)	0.036	0.020	0.556	0.013	0.032	0.063
IPF(25%B2A-S)	0.049	0.030	0.606	0.015	0.044	0.089
IPF(10%B2A-S)	0.066	0.035	0.535	0.021	0.063	0.118
IPF(AFC)	0.065	0.024	0.370	0.036	0.062	0.099
IPF(Null)	0.078	0.026	0.334	0.042	0.081	0.107
All B2A-S	0.164	0.048	0.290	0.103	0.165	0.222
100%B2A-S	0.166	0.054	0.323	0.085	0.168	0.231
75%B2A-S	0.173	0.054	0.310	0.099	0.177	0.241
50%B2A-S	0.190	0.055	0.289	0.115	0.188	0.262
25%B2A-S	0.213	0.064	0.300	0.135	0.212	0.296
10%B2A-S	0.255	0.074	0.290	0.164	0.250	0.353
AFC	0.253	0.104	0.411	0.152	0.220	0.430
Null	0.445	0.124	0.280	0.325	0.434	0.683

The average *HD* values obtained when comparing the various period-level matrices to the corresponding reference matrices are plotted in Figures 2a and 2b. Figures 2a and 2b correspond, respectively, to results obtained when using the *B2A-S* and *IPF(B2A-S)* matrices as references. Similar to what was done when conducting the trip-level analysis, the average values obtained from estimates produced using the IPF method with different sample sizes of the B2A survey data to form the seed (base) matrix and from matrices produced using the sampled survey data directly are plotted as a function of the sample size taken. Once again, the average *HD* value of the *IPF(Null)* B2A flow estimates corresponds to a B2A survey sample size of 0%, and the average *HD* values of the *AFC*, *IPF(AFC)*, and *Null* matrices are plotted as horizontal lines, since these estimates do not depend on the sample size of the B2A survey data.

The general trends in the graphs of the mean *HD* values for period-level B2A flow matrices (Figures 2a and 2b) are similar to those seen in the graphs of mean *HD* values for bus trip-level matrices (Figures 1a and 1b). However, the magnitudes of the mean *HD* values are lower at the period-level, indicating that the errors in estimating bus trip-level matrices are diminished when aggregating to the more practically interesting period-level matrices.

The graphs and tables also show that *IPF(AFC)* estimates perform much better (much lower *HD* values) than the *AFC* estimates regardless of the reference, that the *IPF(X%B2A-S)* estimates perform much better than the *X%B2A-S* estimates for all B2A survey data sample sizes when the *IPF(B2A-S)* is used as the reference, and that the *IPF(X%B2A-S)* estimates perform better than the *X%B2A-S* estimates for B2A survey data sample sizes less than 25% when the *B2A-S* is used as the reference. Since 25% B2A-S sample sizes are considered larger than what would typically be expected in practice, the implication is that using APC data, as is done through the IPF method in this study, can improve B2A flow estimates.

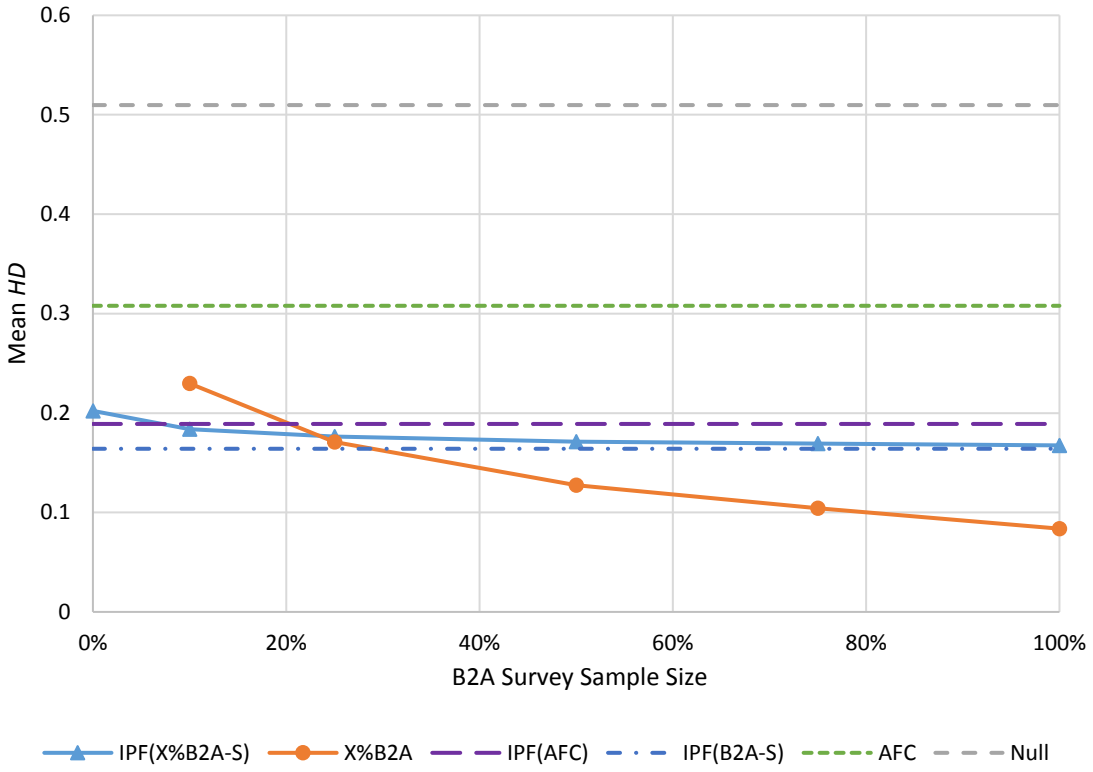

Figure 2a: Mean HD values comparing period B2A matrices determined or estimated from different datasets to reference matrix B2A-S

As expected, the *IPF(X%B2A)* estimates improve (i.e., exhibit lower *HD* values) as the B2A survey sample size increases. However, similar to what was seen with the bus trip-level results, the differences in mean *HD* values obtained when using sample sizes of 100% and those obtained when using sample size of 0% (i.e., *IPF(Null)*) is small compared to the differences obtained when using the uninformative *Null* or *AFC* estimates and when using *IPF(Null)* estimates. The improved accuracy (i.e., reduction in mean *HD* values) associated with using practical sample sizes between 10% and 25% rather than not collecting any B2A survey data (i.e., estimating B2A flows with *IPF(Null)* or *IPF(AFC)* matrices) appears to be very small. That is, most of the B2A information that can be obtained can be obtained by using APC data without the need to collect B2A survey data. Processing available AFC data leads to slightly better estimates than the *IPF(Null)* estimates, when the AFC data are used to form a seed matrix for the IPF method. However, using the AFC data to form B2A estimates without using APC data does not lead to favorable performance.

As was done when investigating the trip-level results, the percentage of times – considering all route-period-directions and B2A survey data samples – that the *IPF(Null), IPF(AFC),* and A*FC* matrices (which can be estimated without any onboard B2A survey data) are closer to the reference than are the *IPF(25%B2A-S)* and *IPF(10%B2A-S)* matrices are determined. Comparisons to *25%B2A-S* and *10%B2A-S* matrices are also determined. The percentages are presented in Table 6.

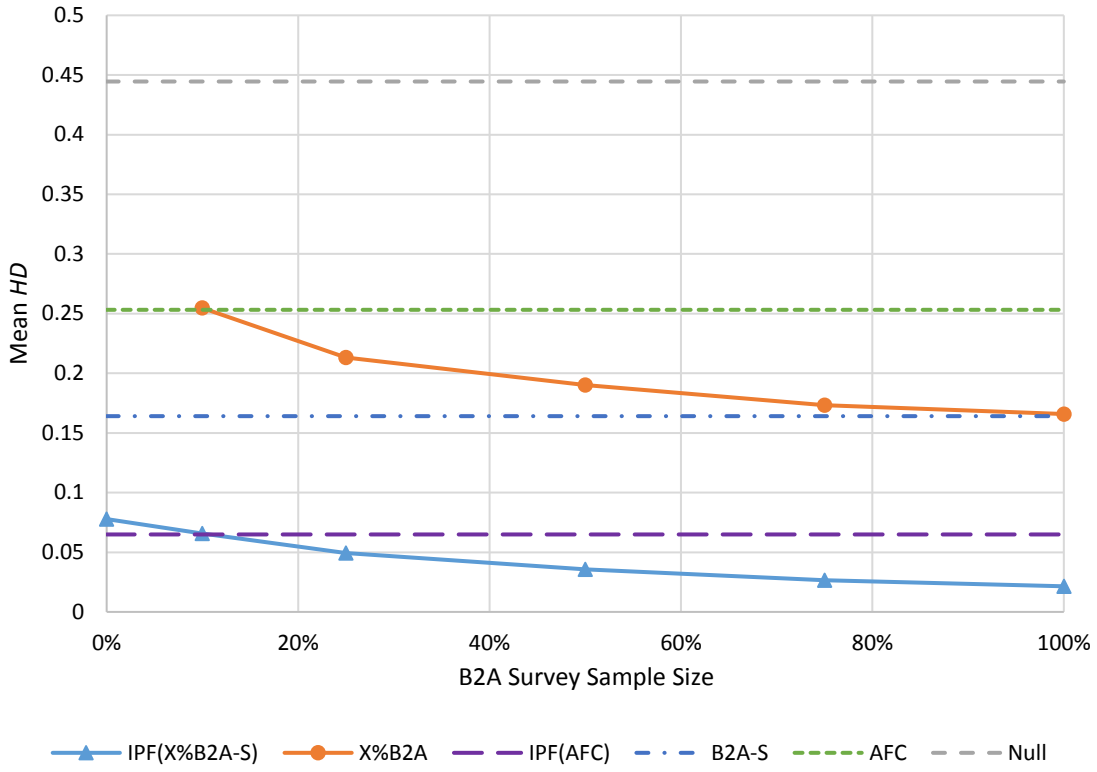

Figure 2b: Mean HD values comparing period B2A matrices determined or estimated from different datasets to reference matrix IPF(B2A-S)

Table 6: Percentage of times IPF(Null), IPF(AFC), and AFC matrices are closer than the 10%B2A-S, 25%B2A-S, IPF(10%B2A-S), and IPF(25%B2A-S) matrices to the reference B2A matrices

	B2A-S Reference			IPF(B2A-S) Reference		
	IPF(Null)	IPF(AFC)	AFC	IPF(Null)	IPF(AFC)	AFC
10%B2A-S	57.5%	65.7%	21.4%	99.6%	100.0%	62.9%
25%B2A-S	33.6%	38.9%	8.2%	100.0%	100.0%	40.4%
IPF(10%B2A-S)	14.3%	31.8%	6.4%	31.1%	51.1%	0.0%
IPF(25%B2A-S)	6.8%	13.9%	6.4%	13.2%	21.4%	0.0%

The percentages associated with *IPF(X%B2A-S)* comparisons are less than those associated with *X%B2A-S* comparisons, supporting the finding that *IPF(X%B2A-S)* matrices are better than *X%B2A-S* matrices for B2A survey data sample sizes between 10% and 25%. Therefore, if B2A survey data are collected, combining these data with available APC data would improve the accuracy of period-level B2A flow estimates. The *IPF(Null)* or *IPF(AFC)* matrices are closer than the *IPF(10%B2A-S)* or *IPF(25%B2A-S)* matrices to the reference less than 50% of the time, except when comparing the *IPF(AFC) and IPF(10%B2A-S)* matrices to the *IPF(B2A-S)* reference. However, the small differences in mean *HD* values (Figures 2a and 2b) and in the percentile values (Tables 5a and 5b) indicate that when the *IPF(X%B2A-S)* matrix is closer than the *IPF(Null)* or *IPF(AFC)* matrix to the reference matrix, it is only by a small amount.

Given that the *IPF(Null)* and *IPF(AFC)* matrices outperform even the *IPF(25%B2A-S)* matrices a non-trivial percentage of times, the results support the implication that little improvement is obtained in estimating period-level B2A flow matrices by collecting B2A survey data, rather than using only available APC and AFC data.

5. Impacts of B2A Flow Estimates on Expanded Characteristics

The investigation of the impacts of the estimated B2A flow matrices on the resulting expansions of the SE&T characteristics collected from onboard personal interview surveys (OBS) is presented in this section. The characteristics collected in the OBS are listed and explained in Appendix D. Of the 58 characteristics, the 34 that reflect socio-economic and travel variables are used in the empirical investigation conducted in this study.

The approach developed and applied to expand the 34 empirically observed SE&T characteristics considered is a rigorous extension of the approach used in AECOM (2014). The concept involves scaling the proportions of responses to questions on SE&T characteristics posed in the OBS by B2A pair to match the proportions obtained in the estimated B2A flow matrix by route, direction, and time-of-day period. The motivation for this expansion is that characteristics could be associated with B2A pair – because individuals with different characteristics travel between different locations – but because of response, administration, or other biases, the proportions of respondents to the OBS by B2A pair may be very different from the proportions of passengers actually travelling on the B2A pair. That is, in this approach, the SE&T characteristics for a responding individual passenger traveling between a specific B2A pair are assumed to be accurate. However, the number of responses from passengers traveling along that B2A pair could be over- or under-sampled in the OBS dataset depending on the characteristic. To correct for such sampling bias, each SE&T characteristic corresponding to each B2A pair is adjusted to reflect the proportion of travelers associated with that B2A pair with respect to the entire traveling population on the route-direction and time-of-day period of interest. This approach is presented in more detail and illustrated by an example in Appendix E.

The OBS data are organized to provide proportions of SE&T characteristics by route-direction-period B2A pair. The expansion by the corresponding B2A proportions would be straightforward if nonzero passenger flows were estimated in the B2A flow matrix for every B2A pair for which OBS responses were recorded and if passengers were recorded in the OBS survey for every B2A pair with a nonzero flow estimate. However, because of B2A survey and AFC sample sizes, APC measurement errors, and biases in the OBS surveys, it is possible – and common – that passenger flows may be zero in a cell of either the B2A or OBS matrix. The approach developed to provide reasonable flow probability estimates where zero flow values were recorded in the B2A matrices is described in Section 3 and Appendix C. To determine the characteristics for a segment-to-segment pair with a zero in the corresponding cell of the OBS matrix, the characteristics associated with passengers boarding at the boarding segment and with passenger alighting at the alighting segment in the pair are adopted. The details of the method and an illustrative example are presented in Appendix F.

To empirically investigate the effect of the estimated B2A flow matrices on the expanded characteristic proportions, the SE&T characteristic data obtained from the OBS for each route, direction, and time-of-day period are expanded using the period-level B2A flow matrix estimates discussed above. These proportions are then compared to reference characteristic proportions determined by expanding the

same OBS SE&T characteristic data using a reference B2A flow matrix. When estimating B2A flow matrices, the *Null* matrix was used as the "uninformative" matrix – that is, the matrix that would be used in the absence of B2A survey, AFC, or APC data (and without any historical estimates of these matrices). If one were to estimate SE&T characteristic proportions without any of these data sources, it would make sense to use the characteristic proportions observed in the OBS data, without any expansion. Therefore, the *Null* in this case refers to these proportions.

The original intention was to compare the characteristic proportions expanded using an estimated B2A flow matrix to characteristic proportions expanded using ground-truth B2A flows. However, the difficulties with the manually collected B2A survey data necessitated the expansion by reference, rather than ground-truth, B2A flow matrices. The *B2A-S* and the *IPF(B2A-S)* flow matrices are used for expanding the OBS SE&T characteristic data to determine the reference SE&T characteristic proportions.

The *HD* metric is again used, this time to measure the difference between a set of SE&T characteristic proportions expanded using an estimated B2A flow matrix and a set of reference characteristic proportions. Analogous to the comparisons of B2A matrices, a set of characteristic proportions associated with a lower *HD* value is considered more similar to the reference set of characteristic proportions than a set with higher *HD* value.

5.1 Attenuation of Errors when Expanding to Characteristics

To begin exploring the effects of the estimated flow matrices on the expanded SE&T characteristic proportions, the relationship between differences in the estimated and reference B2A flow matrices and the corresponding differences in the resulting sets of expanded characteristic proportions is investigated. Figures 2a and 2b present scatter plots of the 99,008 (7 routes x 2 directions x 2 periods x 34 SE&T characteristics x 104 B2A flow estimation methods and multiple samples) pairs of *HD* values (capturing pairs of differences). The scatter plots in Figures 2a and 2b, respectively, are developed using *B2A-S* and *IPF(B2A-S)* as the reference matrices. On the x-axis, the *HD* values are between B2A flow expansion matrices, determined by some method using some sample of the B2A survey dataset, and the corresponding reference (*B2A-S* or *IPF(B2A-S)*) flow matrices. On the y-axis, the *HD* values are between the SE&T expanded characteristic proportions determined by expanding the OBS data with these flow matrices and the reference characteristic proportions determined by expanding the OBS SE&T characteristic data with the corresponding reference flow matrices.

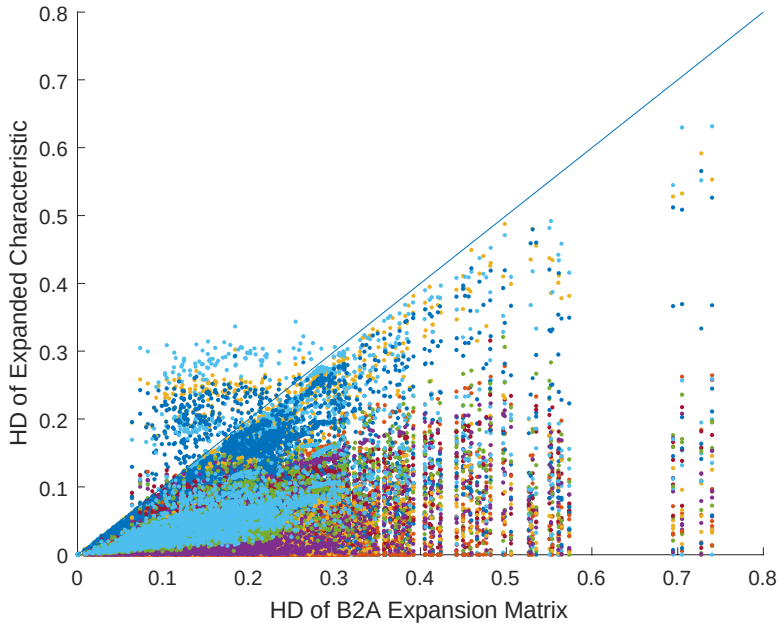

Figure 2a: Scatter plot of HD values of B2A matrices used to expand characteristics and resulting HD values of expanded characteristics using B2A-S matrix as reference

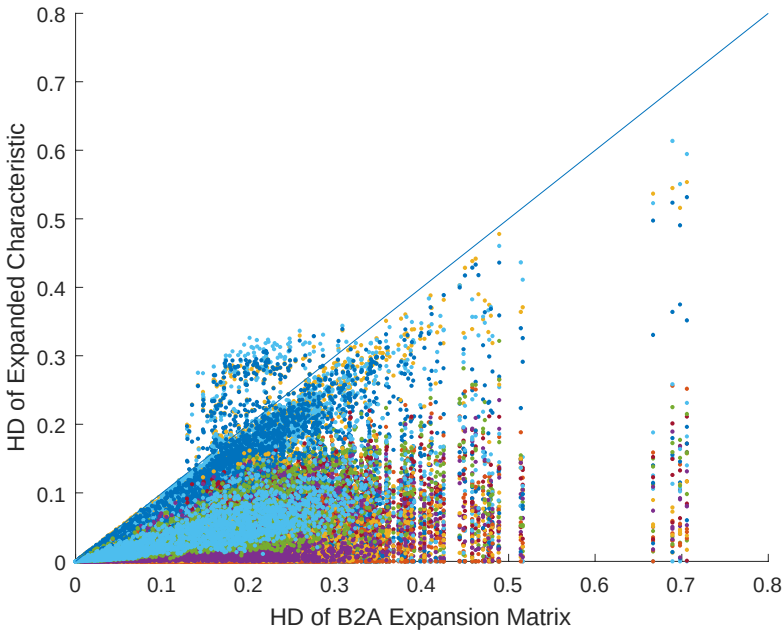

Figure 2b: Scatter plot of HD values of B2A matrices used to expand characteristics and resulting HD values of expanded characteristics using IPF(B2A-S) matrix as reference

The figures show that the vast majority of points (representing pairs of corresponding *HD* values) lie below the line of slope equal to one (the so called "45° line"), indicating that the differences in

expanded characteristic proportions are for the most part less than the differences in the B2A flow inputs. That is, the differences (i.e., "errors" if the reference B2A flow matrix could be considered the ground-truth) in the estimated B2A flow matrices are attenuated in the resulting differences (i.e., "errors" if the SE&T characteristic proportions expanded with the reference B2A flow matrix could be considered to achieve the best corrections to sampling biases) in expanded SE&T characteristic proportions. Indeed, 99.4% and 99.5%, respectively, of the B2A differences are attenuated in this way when considering the *B2A-S* and *IPF(B2A-S)* matrices, respectively, as the reference. An investigation of the results reveals that only "Home Zip", "Origin TAZ", "Destination TAZ", and "Boarding TAZ" had more than 1% of the paired points lying above the line of slope one, where differences in B2A flow matrices are amplified in the resulting differences in the expanded characteristic proportions. (Percentages above the line ranged from 2.1% to 5.6% for these four characteristics.) These characteristics are all highly associated with boarding and alighting stop locations, which are directly associated with the spatial flow patterns represented by the B2A flows used to expand the characteristic data. As a result, it is not surprising that these characteristics do not exhibit as much attenuation as the other characteristics.

5.2 Expanded SE&T Characteristics Comparisons

In Tables 7a and 7b, summary *HD* statistics are presented for the differences between the SE&T characteristic proportions determined by expanding the OBS characteristic data with the various estimated B2A flow matrices and the characteristic proportions determined by expanding the OBS data with the reference B2A flow matrices. In Table 7a, the reference characteristics are those expanded with the B2A matrix. In Table 7b, the reference characteristics are those expanded with the IPF(B2A-S) matrix.

Table 7a: Summary statistics of HD measure comparing SE&T characteristics expanded by period B2A matrices determined or estimated from different datasets to characteristics expanded by reference matrix B2A-S

	Mean	Standard Deviation	Coeff. of Variation	10th Percentile	50th Percentile	90th Percentile
IPF(All B2A-S)	0.058	0.056	0.968	0.002	0.039	0.140
IPF(100%B2A-S)	0.059	0.057	0.967	0.002	0.039	0.142
IPF(75%B2A-S)	0.059	0.057	0.965	0.002	0.039	0.143
IPF(50%B2A-S)	0.060	0.058	0.963	0.002	0.040	0.144
IPF(25%B2A-S)	0.061	0.058	0.960	0.002	0.041	0.146
IPF(10%B2A-S)	0.062	0.060	0.965	0.002	0.042	0.152
IPF(AFC)	0.064	0.061	0.954	0.003	0.042	0.154
IPF(Null)	0.068	0.065	0.955	0.003	0.046	0.167
100%B2A-S	0.040	0.057	1.446	0.001	0.019	0.092
75%B2A-S	0.046	0.060	1.288	0.001	0.026	0.109
50%B2A-S	0.054	0.062	1.146	0.001	0.033	0.133
25%B2A-S	0.068	0.072	1.061	0.002	0.043	0.173
10%B2A-S	0.084	0.085	1.012	0.002	0.054	0.215
AFC	0.109	0.107	0.989	0.003	0.073	0.253
Null	0.198	0.186	0.938	0.008	0.128	0.492

Table 7b: Summary statistics of HD measure comparing SE&T characteristics expanded by period B2A matrices determined or estimated from different datasets to characteristics expanded by reference matrix IPF(B2A-S)

	Mean	Standard Deviation	Coeff. of Variation	10th Percentile	50th Percentile	90th Percentile
IPF(100%B2A-S)	0.006	0.007	1.257	0.000	0.003	0.016
IPF(75%B2A-S)	0.007	0.008	1.238	0.000	0.003	0.018
IPF(50%B2A-S)	0.009	0.012	1.286	0.000	0.004	0.025
IPF(25%B2A-S)	0.013	0.017	1.319	0.000	0.006	0.034
IPF(10%B2A-S)	0.017	0.021	1.267	0.000	0.008	0.047
IPF(AFC)	0.015	0.017	1.117	0.000	0.009	0.038
IPF(Null)	0.020	0.022	1.090	0.001	0.012	0.053
All B2A-S	0.058	0.056	0.968	0.002	0.039	0.140
100%B2A-S	0.064	0.067	1.045	0.002	0.040	0.168
75%B2A-S	0.067	0.068	1.024	0.002	0.043	0.168
50%B2A-S	0.073	0.071	0.983	0.003	0.047	0.181
25%B2A-S	0.080	0.078	0.978	0.003	0.052	0.200
10%B2A-S	0.091	0.087	0.956	0.003	0.061	0.226
AFC	0.088	0.090	1.018	0.003	0.058	0.211
Null	0.184	0.176	0.957	0.007	0.184	0.462

The mean *HD* values obtained when comparing the SE&T characteristic data expanded with the various period-level B2A flow matrices to the reference characteristic proportions determined by expanding the OBS SE&T characteristic data with the *B2A-S* and *IPF(B2A-S)* flow matrices are plotted, respectively, in Figures 4a and 4b. The plots are shown as a function of the sample size of the B2A survey data forming the B2A flow matrices used for expanding the characteristic data and the seed matrices for the IPF-estimated expansion B2A flow matrices. Similar to the mean *HD* plots for estimated period-level B2A flow matrices shown in Figures 2a and 2b, the mean *HD* values associated with characteristic proportions obtained when expanding by the *IPF(Null)* flow matrices *c*orresponds to a B2A survey sample size of 0%. In addition, the values associated with characteristic proportions obtained when expanding by the *AFC*, *IPF(AFC)*, *IPF(B2A-S)*, *and B2A-S* flow matrices – or when not expanding at all (*Null)* – are plotted as horizontal lines, since they do not depend on the sample size of the B2A survey data.

The *HD* values in Tables 7a and 7b and Figures 4a and 4b are markedly lower than those in Tables 5a and 5b and Figures 2a and 2b, supporting that "errors" in estimating B2A flows are attenuated when considering their effects on SE&T characteristic expansion.

The results in Tables 7a and 7b and Figures 4a and 4b also indicate that, regardless of the reference, the SE&T characteristic proportions expanded using the *IPF(Null)* or *IPF(AFC)* matrices are markedly better (i.e., markedly closer to the reference characteristic proportions) than the characteristic proportions expanded using only the AFC data and the proportions contained in the OBS data (i.e., the *Null*). The characteristic proportions expanded with the *X%B2A-S* matrices are better than the proportions expanded with the *IPF(X%B2A-S)* or the *IPF(AFC)* matrices only when B2A survey sample sizes are greater than 40% and only when considering the characteristics expanded by *B2A-S* matrices as a

reference. Since 40% B2A survey data sample sizes are much larger than those that would be expected in practice, it appears that one should use APC data in estimating B2A flow matrices for expansion.

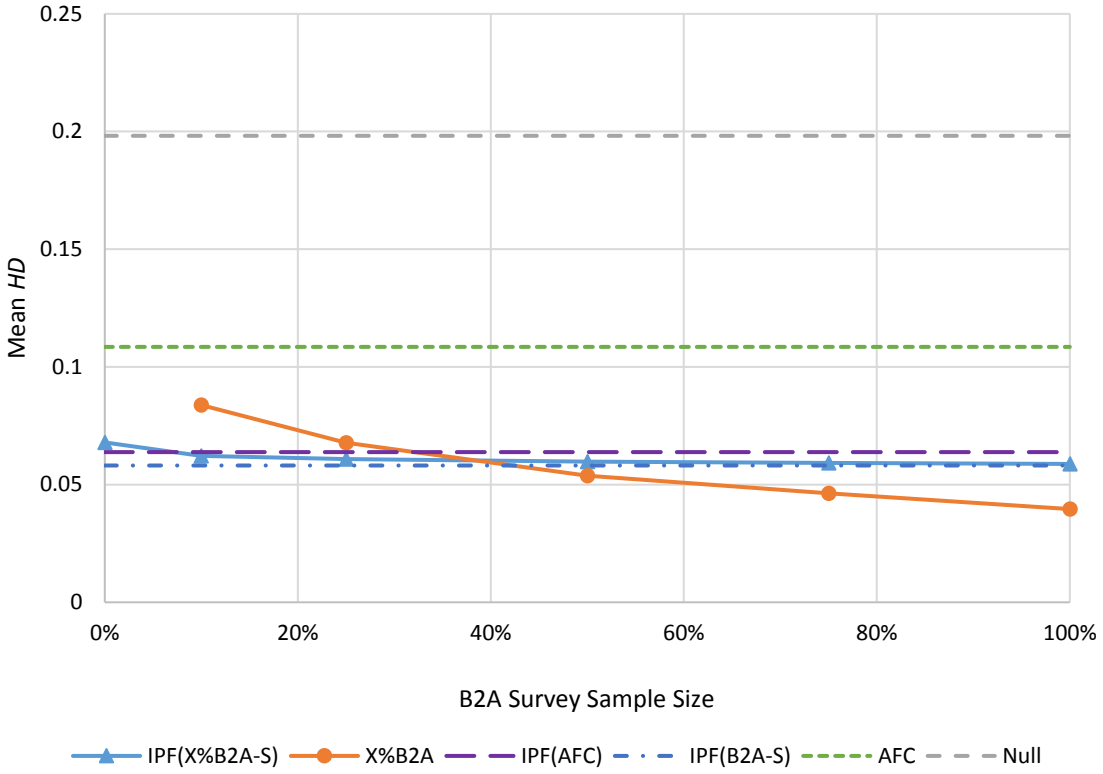

Figure 4a: Mean HD values comparing SE&T characteristics expanded by period B2A matrices determined or estimated from different datasets to characteristics expanded by reference matrix B2A-S

The characteristic proportions expanded with the *IPF(X%B2A-S)* matrices improve (i.e., have lower *HD* values) as the B2A sample size increases. However, similar to what was seen when investigating the improvement to the B2A flow matrices at both the bus trip- and period-levels, the improvements in the *IPF(X% B2A-S)*-expanded characteristics associated with increases in the B2A survey sample size are small when compared to the improvements associated with expanding by the *IPF(Null)* or *IPF(AFC)* matrices rather than the *AFC* matrices or not expanding at all (i.e., the *Null)*. When considering the 10% to 25% range of B2A survey sample sizes (bracketing sample sizes expected in practice), one sees that, regardless of the reference, the mean *HD* values associated with *IPF(X%B2A-S)*-expanded characteristics are only slightly less than those associated with the 0% B2A survey sample size *IPF(Null)*-expanded characteristics. Even less, if any, reduction in mean *HD* values is offered when expanding by the *IPF(X%B2A-S)*, rather than the *IPF(AFC)* matrices. Similar to what was seen when considering B2A matrix estimation, when considering the effect of the estimated B2A matrix on the expanded characteristics, the implication is that either the *IPF(Null)* or *IPF(AFC)* matrices (which can be obtained without the expense associated with B2A surveys) can be used for expansion with very little or no loss in accuracy, compared to expanding by matrices estimated with B2A survey data collected with sample sizes typical in practice. On the other hand, expanding by the *AFC* matrices, which also require no B2A survey data, performs appreciably worse than expanding by the *IPF(X%B2A-S)* estimates that use B2A survey data. As when considering the effect only on B2A estimation, when considering the effect on SE&T characteristic

expansion, the implication is that using APC data is useful in determining the B2A matrices used for expansion and that incorporating B2A survey data offers little improvement.

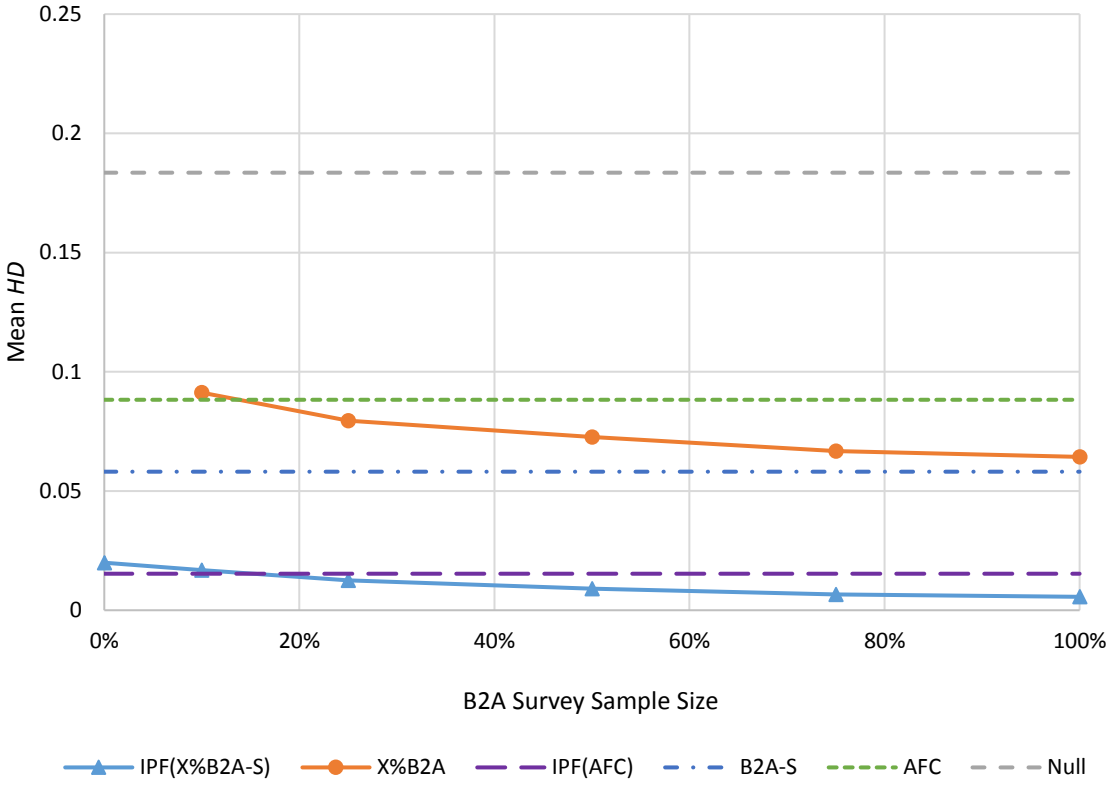

Figure 4b: Mean HD values comparing SE&T characteristics expanded by period B2A matrices determined or estimated from different datasets to characteristics expanded by reference matrix IPF(B2A-S)

As was done when investigating the accuracies of the period-level B2A flow estimates, the percentage of times (across routes, periods, directions, and samples) that the SE&T characteristic proportions expanded with the *IPF(Null)*, *IPF(AFC)*, and *AFC* B2A flow matrices are closer to the reference characteristic proportions (i.e., have lower *HD* values) than the characteristic proportions expanded with the *10%B2A*, *25%B2A*, *IPF(10%B2A)*, and *IPF(25%B2A)* matrices are determined. These percentages are presented in Table 8.

Table 8: Percentage of times SE&T characteristics expanded with IPF(Null), IPF(AFC), and AFC flow matrices are closer than characteristics expanded with the 10% B2A-S, 25% B2A-S, IPF(10% B2A-S), and IPF(25% B2A-S) flow matrices to reference characteristics

	B2A-S Reference			IPF(B2A-S) Reference		
	IPF(Null)	IPF(AFC)	AFC	IPF(Null)	IPF(AFC)	AFC
10%B2A-S	55.9%	60.0%	35.0%	94.2%	95.8%	53.8%
25%B2A-S	44.1%	46.9%	28.4%	93.6%	95.0%	46.0%
IPF(10%B2A-S)	27.9%	41.4%	21.7%	36.1%	50.3%	4.4%
IPF(25%B2A-S)	25.2%	36.3%	21.1%	23.8%	33.4%	2.9%

As with the period-level B2A flow matrix analysis, the percentages are much lower when comparing to characteristics expanded with *IPF(X%B2A-S)* matrices than when comparing to characteristics expanded with *X%B2A-S* matrices, supporting the finding that the *IPF(X%B2A-S)*-expanded characteristics are better than the *X%B2A-S*-expanded characteristics. Therefore, if B2A survey data are to be collected, it is advisable to also use already available APC data when estimating B2A matrices for expansion of SE&T characteristics.

The percentage obtained when comparing *IPF(AFC)*-expanded characteristics to *IPF(10%B2A*-S)-expanded characteristics when using *IPF(B2A-S)*-expanded characteristics as a reference is approximately the same in Table 8 (50.3%) as in Table 6 (51.1%). The other percentages obtained when making comparisons to *IPF(X% B2A-S)*-expanded characteristics are markedly greater in Table 8. As discussed above, the overall reduction in the *HD* values associated with SE&T characteristics from the values associated with period-level B2A flow matrices demonstrates an attenuation in "errors" when using the estimated B2A flow matrices to expand SE&T characteristics. The attenuation occurred when considering expansion by all B2A estimation methods and data sets. The increased percentages seen in Table 8 indicate that the attenuation has a greater effect on the *IPF(Null)-*, *IPF(AFC)-*, and *AFC*-expanded characteristics than on the *IPF(X%B2A-S)*-expanded characteristics. Nevertheless, most of the percentages associated with *IPF(Null)-* and *IPF(AFC)*-expanded characteristics are still less than 50%, indicating that the *IPF(X%B2A-S)*-expanded characteristics are closer to the reference characteristics. However, the small differences in the *HD* values between these sets of characteristics seen in Tables 7a and 7b and Figures 4a and 4b indicate that any improvements are small. Given the fairly large percentages in Table 8, the results support the implication that little improvement in expanded characteristics is obtained by collecting B2A survey data for use in estimating the B2A flow matrices used in expansion. Rather, estimating the B2A flow matrices using APC data to produce *IPF(Null)* and, especially, *IPF(AFC)* matrices would appear sufficient for expansion purposes.

6. Conclusions

B2A flows are estimated in multiple ways using B2A survey, APC, and AFC data collected on COTA routes, and the estimates are compared to reference B2A flows considered to represent the ground-truth most accurately. In addition, passenger SE&T characteristic proportions expanded using the multiple B2A flow estimates are compared to reference characteristic proportions considered to represent sampling bias corrections most accurately. The results lead to three main conclusions. First, expensive and labor-intensive B2A survey data collected at typical sample sizes offer little improvement in the accuracy of B2A flows at the bus trip- and time-of-day period-levels, compared to estimates that rely only on readily available APC and AFC data. Second, the "errors" in the estimated period-level B2A flow matrices used to expand the SE&T data are attenuated in the resulting "errors" in the expanded SE&T characteristic proportions. Third, B2A survey data collected at sample sizes typical of those collected in practice offer little, if any, improvement in expanding characteristic data compared to using B2A flow matrices estimated from readily available APC data and, especially, from a combination of APC and AFC data.

In this study, the state-of-the-practice IPF method was used to estimate B2A flows from APC data and from APC and AFC data. Recent advances in the state-of-the-art of estimating B2A flows from boarding and alighting count data have shown the potential to provide estimates that are superior to those provided by the IPF method when sufficiently large datasets are available (Ji et al., 2014; Ji et al., 2015).

With the increasing adoption of automatic data collection systems by the transit industry, such large datasets are becoming increasingly available, and these methods are likely to become more commonly adopted. Investigating these methods in the context of the empirical investigations conducted in this study would be worthwhile.

In addition, it is noted that the AFC data available in this study were subject to AVL errors that rendered the locations of stops associated with AFC transactions unusable. As a result, stop locations had to be corrected through a process involving the matching of the AFC dataset with the APC dataset based on the reliable timestamps in both datasets. Given that only around 20% of the COTA fleet is equipped with APC technology and the nature of inferring alighting stops from AFC transactions, only a very small percentage of AFC passenger transactions led to usable AFC B2A flows. Given the impressive performance that resulted when combining even this small percentage of AFC passenger transactions with APC data, it would be enlightening to reinvestigate the contributions of using AFC data, either alone or in combination with APC data, when alighting stops can be inferred from AFC technologies for a larger percentage of the ridership.

References

AECOM, 2014. 2013 COTA On-Board Survey Expansion. Memorandum to Mr. Michael McCann, COTA from Mr. Tim Palermo, Ms. Jamie Snow, Mr. David Schmitt, AECOM dated February 6, 2014.

Ben-Akiva, M., Macke, P., and Hsu, P., 1985. Alternative methods to estimate route-level trip tables and expand on-board surveys. Transportation Research Record, 1037, 1-11.

Campus Transit Lab (CTL), The Ohio State University, 2015. http://transitlab.osu.edu/campus-transit-lab.

Furth, P.G., Strathman, J.G., Hemily, B., 2005. Making automatic passenger counts mainstream: Accuracy, balancing algorithms, and data structures. Transportation Research Record, 1927, 207-216.

Ji, Y., Mishalani, R.G., McCord, M.R., Goel, P., 2011. Identifying Homogeneous Periods in Bus Route Origin-Destination Passenger Flow Patterns from Automatic Passenger Counter Data. Transportation Research Record, 2216, 42-50.

Ji, Y., Mishalani, R.G., McCord, M.R., 2014. Estimating Transit Route OD Flow Matrices from APC Data on Multiple Bus Trips Using the IPF Method with an Iteratively Improved Base: Method and Empirical Evaluation. Journal of Transportation Engineering, 140(5), 040140081-040140088.

Ji, Y., Mishalani, R.G., McCord, M.R., 2015. Transit Passenger Origin-Destination Flow Estimation: Efficiently Combining Onboard Survey and Large Automatic Passenger Count Datasets. Transportation Research Part C: Emerging Technologies (Special Issue: Big Data in Transportation and Traffic Engineering), 58, 178-192.

McCord, M. R., Mishalani, R. G., Goel, P., and Strohl, B., 2010. Iterative proportional fitting procedure to determine bus route passenger origin-destination flows. Transportation Research Record, 2145, 59-65.

McCord, M.R., Mishalani, R.G., Hu, X., 2012. Bus Stop Grouping for Aggregation of Route level Passenger Origin-Destination Flow Matrices. Transportation Research Record, 2277, 38-48.

Mishalani, R.G., Ji, Y., McCord, M.R., 2011. Empirical Evaluation of the Effect of Onboard Survey Sample Size on Transit Bus Route Passenger OD Flow Matrix Estimation Using APC Data. Transportation Research Record, 2246, 64-73.

Simon, J., and Furth, P. G., 1985. Generating a bus route o-d matrix from on-off data. Journal of Transportation Engineering, 111(6), 583-593.

Appendix A: Screening of B2A-S Data

To compare bus trip-level estimates, criteria were developed and implemented to screen out bus trips with what appeared to be particularly poor quality B2A survey data.

Two metrics were developed based on the empirical data, and thresholds were set to screen out bus trips with metric values that did not satisfy the threshold conditions. One metric was the ratio *R* of trip volume obtained from the B2A survey to the trip volume obtained from the APC data after applying the APC balancing procedure discussed in Section 2:

$$R_i = \frac{\text{Passenger volume on bus trip } i \text{ in B2A-S data}}{\text{Passenger volume on bus trip } i \text{ in APC data}} \quad \text{(A-1)}$$

It was observed, and confirmed by discussion, that the data collectors often obtained data on a fairly small number of passengers who travelled on the bus trip. It was, therefore, believed that the representativeness of the *B2A-S* flow matrix for a trip would be best for *R* values approximately equal to 1. Trips were to be considered in the bus trip level analysis only if the *R* value for the trip fell between lower bound *a* and upper bound *b* thresholds:

$$a \leq R \leq b \quad \text{(A-2)}$$

The determination of the thresholds was based on the assumption that the *IPF(Null)* matrix, which did not use any *B2A-S* data, would reflect the spatial B2A flows with some validity. Therefore, bus trips with better quality B2A survey data would generally have *B2A-S* matrices that are "closer" to the *IPF(Null)* matrices than bus trips with poorer quality B2A survey data.

Therefore, for each bus trip, the *R* value (indicating the degree of equality between B2A survey and APC volumes for the trip) and the *HD* value of Eq. (A-1), (indicating the difference in the trip's *B2A-S* and *IPF(Null)* matrices, which were derived from the B2A survey and APC data, respectively) was determined. A scatter plot of the *(R, HD)* values was developed for the 225 trips for which B2A survey data for the entire trip (i.e., trips that did not suffer from the "B2A data partial trip" difficulty discussed in Section 2 and Appendix B) and APC data that passed the APC filtering tests described in Section 2 were available. To reduce the "noise" In the plot, moving averages were taken. Specifically the trips were arranged by increasing *R* value. Then, the average *R* and average *HD* values were calculated for trips 1 through 20 (the 20 trips with lowest *R* value). Then, *R* and *HD* averages were taken for trips 2 through 21. The process was continued, each time increasing the beginning of the "20 trip window" by 1, until averages were computed for the 20 trips with highest *R* values. A scatter plot of these averages is presented in Figure A-1.

The plot shows a general decrease in average *HD* value as average *R* increases to a value of 1.0, with some local variations in the decreasing pattern This decreasing pattern is consistent with the hypotheses leading to the use of the *R* metric, namely, that: (i) the B2A-S data would be more representative of true passenger B2A flows on the bus trip as the number of passenger for whom B2A survey data were collected became closer to the total passenger volume on the trip, (ii) that the APC volume was representative of the total volume on the trip, and (iii) the *IPF(Null)* matrix was generally representative of passenger flows on the trip.

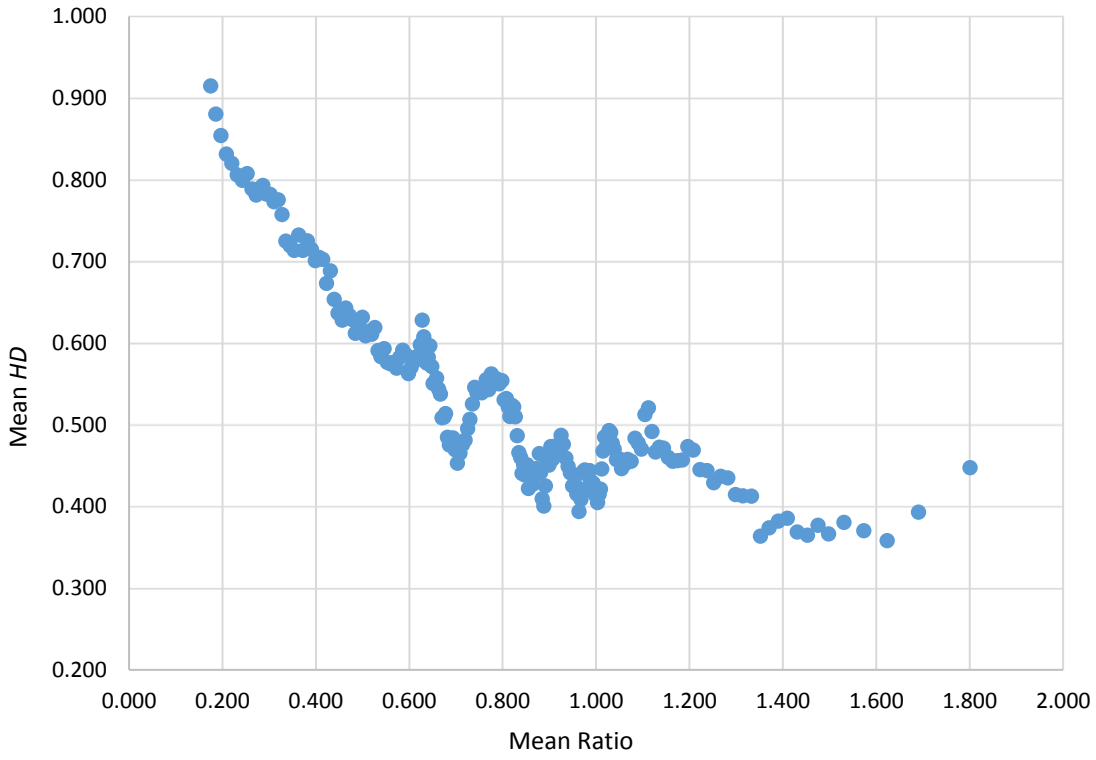

Figure A-1: Moving averages of Mean *HD* between *B2A-S* and *IPF(Null)* bus trip matrices and Mean Ratio *R* of B2A survey and *APC* trip volumes; averages taken over 20 consecutive trips

Because of errors in the APC data, It would be possible for B2A survey volumes to exceed APC volumes on some trips and still lead to both *B2A-S* and *IPF(Null)* matrices that are representative of passenger flows on the trips. However, the decreasing pattern in the graph after an average *R* value of approximately 1.1 and the very low average *HD* values at average *R* values around 1.4 are surprising. Considerable effort was devoted to error checking and attempting to explain the result. No errors were discovered, and no plausible explanation was developed. Because it seemed implausible to have ratios much greater than 1.0 (Indicating errors in B2A-S data, APC data, or both), and because of the local minimum in the average *HD* value at 1.0, it was decided to set the upper bound *b* on *R* at 1.0, that is, to eliminate trips with R values greater than 1.0 from consideration in the trip level comparisons.

The pattern in Figure A-1 indicates that a greater value of lower bound *a* would generally be associated with better correspondence of *B2A-S* and *IPF(Null)* matrices, which would arguably indicate better quality B2A survey data. However, greater values of the lower bound would eliminate more trips and reduce the ability to conduct meaningful comparisons. To investigate the trade-off between quality of the B2A survey data and the number of trips, the average *HD* values (between *B2A-S* and *IPF(Null)* matrices) for all trips with *R* value between trial values of lower bound *a* and the selected upper bound *b* = 1 were investigated. A graph of mean *HD* value for bus trips with *R* between values of *a* = 0, 0.1, 0.2, …, 0.9 and *b* = 1.0 versus number of bus trips with *R* between the *a* and *b* values appears in Figure A-2. Lower values of lower bound *a* lead to more trips in the interval. Therefore, the points in the figure correspond to values of *a* decreasing from 0.9 to 0.0 as one reads from left to right. The orange triangle

in the figure is used to indicate the mean *HD* value and number of trips when considering all 225 trips (including those with *R* greater than one) originally considered.

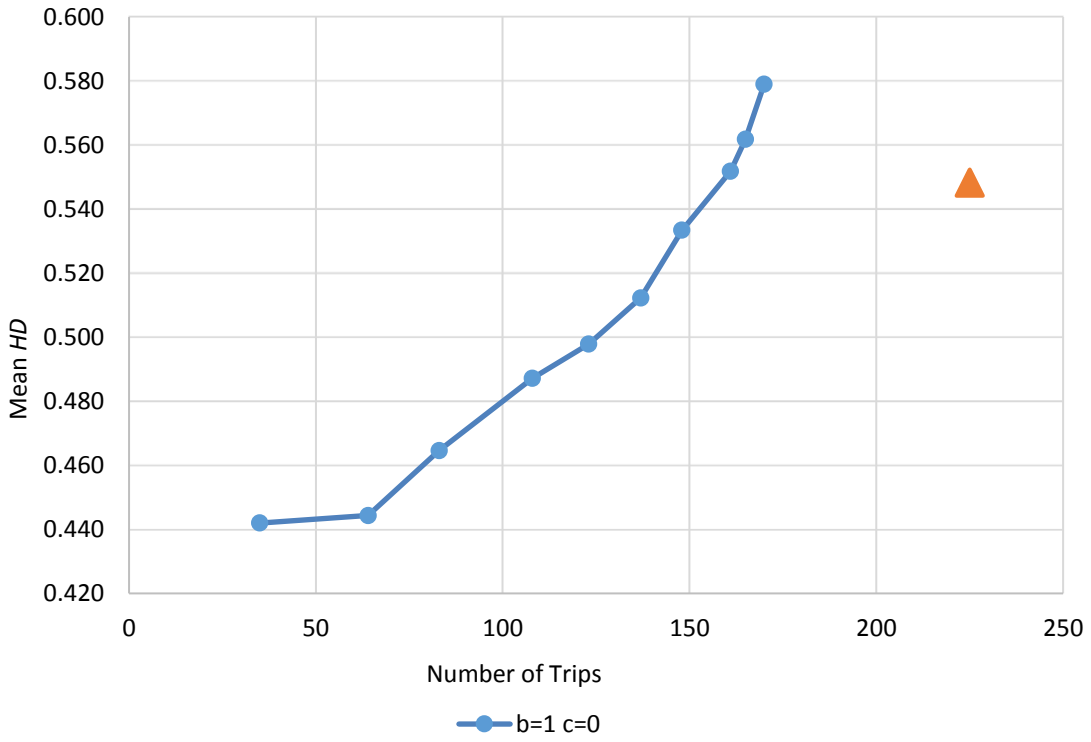

Figure A-2: Mean *HD* between *B2A-S* and *IPF(Null)* matrices for trips with ratio *R* of B2A survey and APC trip volumes between varying lower bound *a* and upper bound *b = 1* versus Number of Trips with R between *a* and *b*; points represent values of a = 0.9, 0.8, 0.7, …, 0.0 moving from left to right

The drastic change in the slope of the graph when moving from the second point (*a* = 0.8) to the third point (*a* = 0.7) indicates that a relatively large increase in average *HD* value (arguably reflecting poorer quality in the B2A survey data) for a relatively small increase in number of trips to be retained in the analysis. Therefore, a lower bound *a = 0.8* on *R* was chosen, resulting in only the 64 trips with *R* value between 0.8 and 1.0 being considered in the trip level investigation of B2A estimation.

A second metric was considered to further screen the 64 trips with *R* value between 0.8 and 1.0. This metric was based on the desire to see boarding "activity" (one or more passengers boarding) in the APC data for a segment when the B2A survey recorded boarding activity (one or more passengers boarding) at the segment, and similarly for alighting activity, The metric P(APC|B2A-S):

$$P(APC|B2A\text{-}s)_i = \frac{\#\ of\ segments\ on\ bus\ trip\ i\ with\ both\ APC\ and\ B2A\text{-}S\ boarding\ (alighting)\ activity}{\#\ of\ segments\ on\ bus\ trip\ i\ with\ both\ B2A\text{-}S\ boarding\ (alighting)\ activity} \quad \text{(A-3)}$$

for a bus trip was based on the concept of conditional probability (specifically, the probability that for a randomly chosen segment on the bus trip, APC boarding (alighting) activity would be seen in the APC

data if boarding (alighting) activity was seen in in the B2A survey data. Values of P(APC|B2A-S) are bounded from above by 1.0. A lower bound *c* was to be chosen so that only trips with:

$$P(APC|B2A\text{-}S)_i \geq c \quad \text{(A-4)}$$

for both boarding *and* alighting activity would be retained for the trip-level B2A estimation investigation.

To choose the value of *c*, the number of trips with $0.8 \leq R \leq 1.0$, the values of P(APC|B2A-S) greater than or equal to $c = 0.\ 0.05, 0.10, 0.15, \ldots, 1.0$, and the average *HD* value between the *B2A-S* and *IPF(Null)* matrices of these trips were determined. The number of trips remained constant for four ranges of the discrete values of investigated – namely, (i) $c = 0.00, 0.05, 0.10, \ldots, 0.65$; (ii) $c = 0.70, 0.75$; (iii) $c = 0.80$, and (iv) $c = 0.85, 0.90, 0.95, 1.0$. The average *HD* value versus the number of trips when considering these thresholds are plotted in Figure A-3.

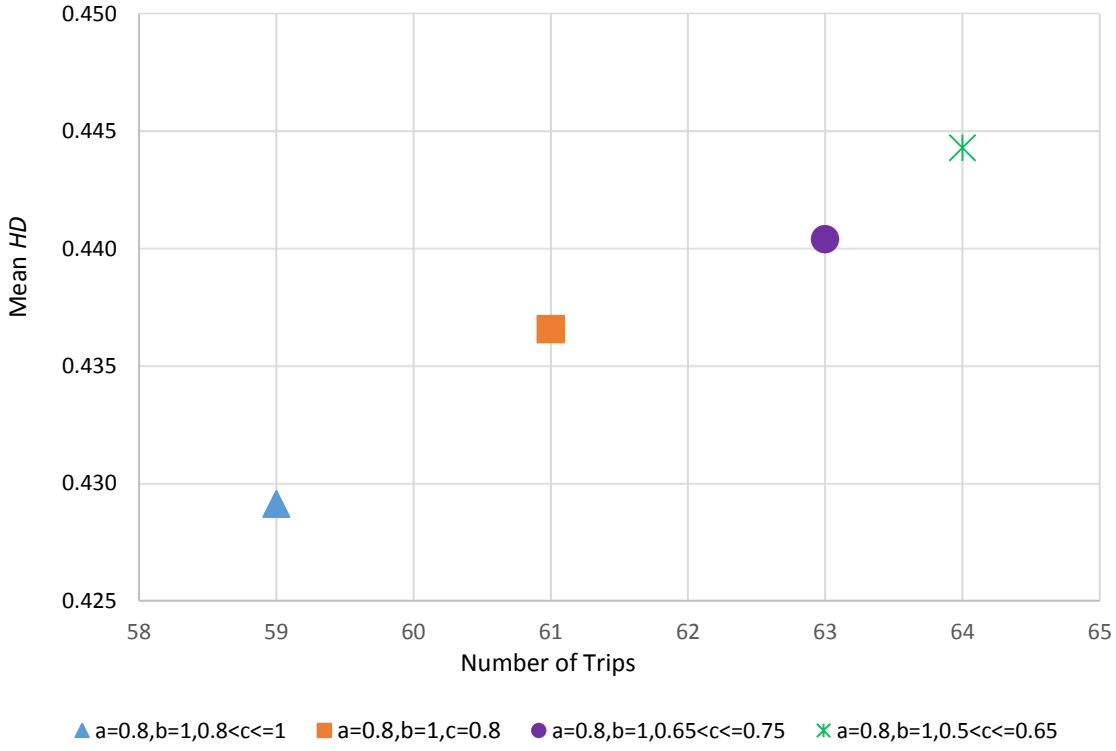

Figure A-3: Mean *HD* between *B2A-S* and *IPF(Null)* matrices for trips with ratio *R* of B2A survey and APC trip volumes between $a = 0.8$ and $b = 1.0$ and B2A-S and APC boarding and alighting activity metric *P(APC|B2A-S)* greater than values of threshold *c* versus Number of Trips satisfying the conditions

Figure A-3 indicates that the average *HD* value decreases with increasing value of *c* (which is associated with decreasing number of trips with $P(APC|B2A\text{-}S)_i \geq c$). As with the result obtained when using the *R* metric above, this result is consistent with the hypotheses that the *P(APC|B2A-S)* metric is reflecting the quality of the B2A survey data on the trip and that the *IPF(Null)* matrices are indicating trip level passenger flows with some validity.

Considering c = 1 (the most stringent level on the quality of the B2A survey date) leads to a reduction of only five trips compared to considering $c = 0$ (the least stringent level on the quality of the B2A survey data). Therefore, a value of c = 1 is selected as the threshold on the *P(APC|B2A-S)* metric.

The threshold $c = 1$ on the *P(APC|B2A-S)* = 1 metric was selected after choosing the thresholds a and b on the volume ratio R. That is, the analyses conducted to choose thresholds a and b were performed with all the trips, which is equivalent to $c = 0$. Therefore, the analyses leading to Figures A-1 and A-2 for trips with c = 0 were repeated for trips with c. The results are presented in Figures A-4 and A-5, respectively. The similarity in the results are reassuring, and the 59 trips with $0.8 < R < 1.0$ and P(APC|B2A-S) ≥ 1.0 were considered as having B2A survey data of sufficient quality to be used in the trip-level investigation of the quality of B2A estimation.

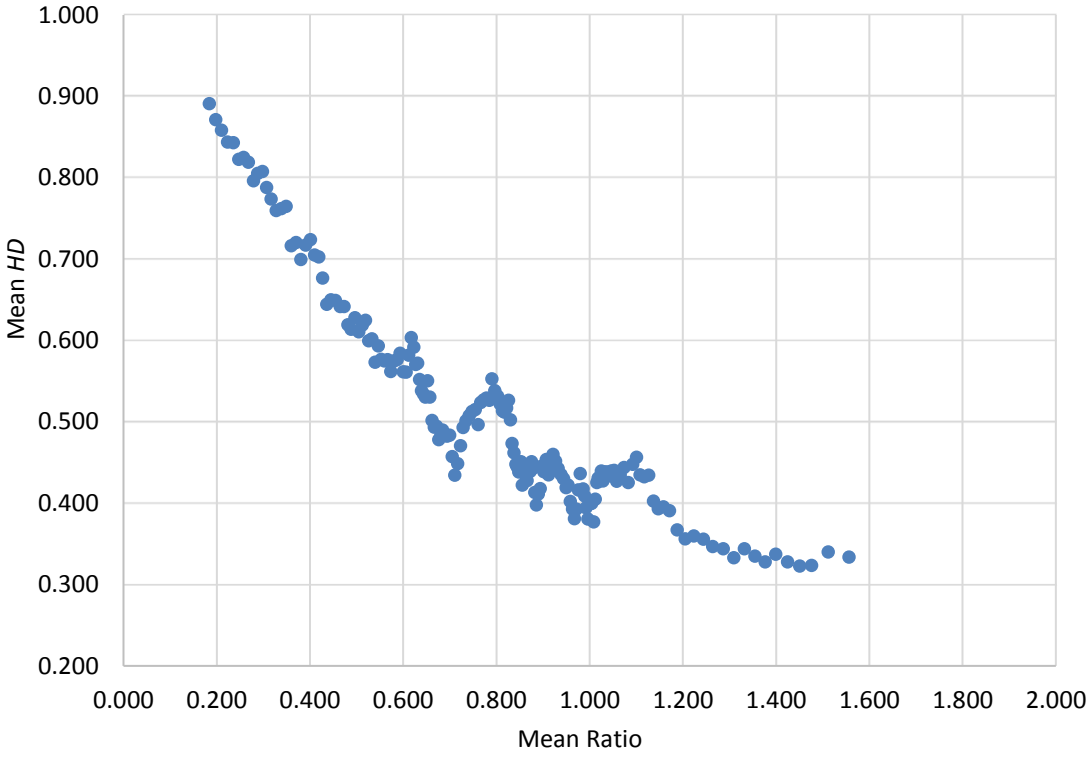

Figure A-4: Moving averages of Figure A-1 recalculated using trips with threshold c on P(APC|B2A-S) metric set to 1, rather than 0

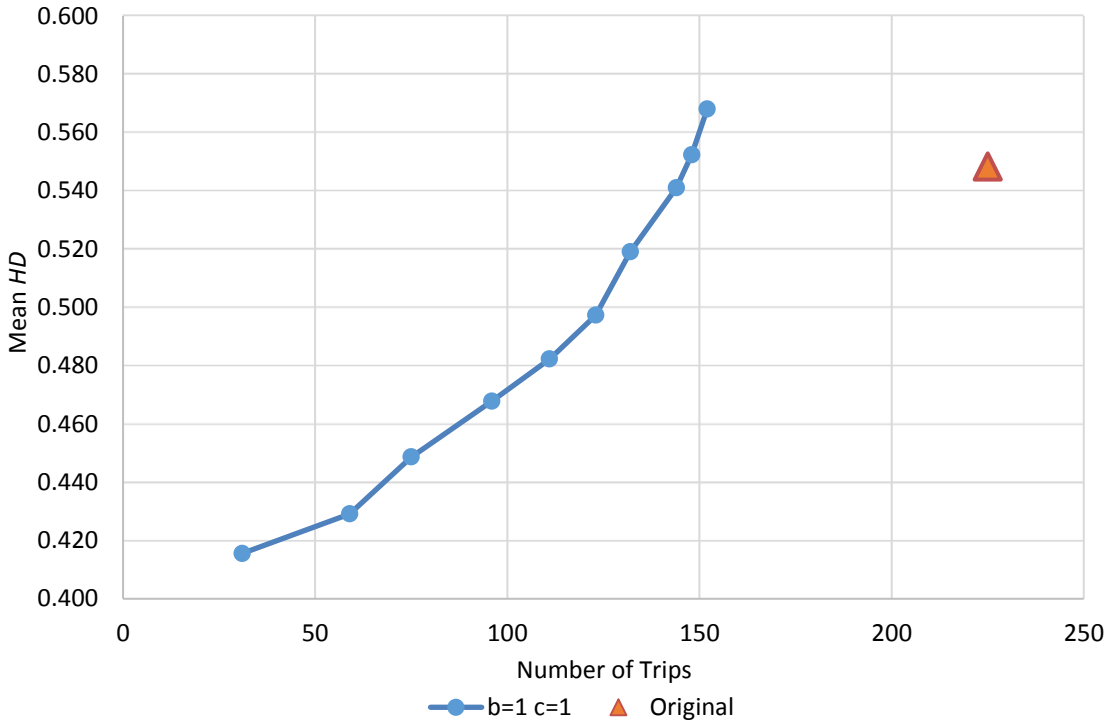

Figure A-5: Mean *HD* versus number of trips remaining analysis represented in Figure A-2 re-plotted after considering threshold *c* on P(APC|B2A-S) metric set to 1, rather than 0

Appendix B: Treatment of Partial Trips

As discussed in Section 2 of the report, when data collectors start and end the B2A data collection at locations other than route terminals, the resulting B2A data will not span entire bus trips. Such a situation leads to an incomplete representation of the passenger flow patterns where only partial bus trips are represented. Given the large number of such B2A data partial trips (56% of the total bus trips observed for the route-directions and periods considered for the period-level B2A flow and SE&T characteristic analyses), it is important to take advantage of these incomplete data. To do so, a method was developed to complete B2A partial trip data based on information available in the complete B2A trip data for a given route-direction and time-of-day period. In the case of partial trips, when data collectors are not onboard for an entire trip, they board or alight at the same stop leading to partial trips where the passenger flows on either end of a trip are observed. Thus, there are two types of partial trips, one type where the upstream end is missing (i.e., data collectors did not start data collection at a terminal) and another type were the downstream end is missing (i.e., data collectors did not finish data collection at a terminal). For a given route-direction and time-of-day period and for the bus trips where the passengers flows related to the upstream end are missing, proportions of passenger flows are determined for the missing portions of the B2A data by considering all the observations on the complete trips and the partial trips where data related to the upstream end are available (i.e., the downstream end is missing). These probabilities and the observed passenger volumes are then used to complete the partial trips where the passenger flows associated with the upstream end are missing. A similar process is followed for bus trips where the passenger flows related to the downstream end are missing.

Appendix C: Replacing Nostructural Zeroes in B2A Passenger Flow Matrices

The relatively small sample sizes in the B2A survey and the inability to infer alighting stops for many of the observed AFC boardings can lead to some segment-to-segment pairs with zero flow in the period-level *B2A-S, X%B2A-S,* and *AFC* matrices. Having such "nonstructural zeroes" in the B2A flow matrices is problematic for the expansion procedure if there are non-zero entries in the corresponding OBS segment-to-segment pair obtained from the personal interview survey. The expansion procedure discussed in Section 5 scales the OBS characteristic data by the B2A flow data. Scaling by a B2A flow entry of 0 will lead to a corresponding value of 0 in the OBS data, which is inconsistent with having observed OBS data for this segment-to-segment pair. In addition, a non-structural zero in the cell of a seed matrix to the IPF procedure will lead to an output value of 0~~zero~~ for that cell. Therefore, the *IPF(B2A-S), IPF(X%B2A-S),* and *IPF(AFC)* matrices are also susceptible to having some cells with flow entries of 0, and these matrices would not be useable for expanding SE&T characteristics.

Because of this difficulty, a procedure was developed to replace values of 0 in a B2A passenger flow matrix with "reasonable" nonzero values. In general, the procedure was designed to replace flow probability values of 0 in an "original" matrix with the greatest flow probabilities in a new probability matrix such that the likelihood of observing the passenger flow volumes associated with the original matrix is still sufficiently large when considering the this new probability matrix.

Specifically, a "combined" probability matrix **C** is formed as a weighted combination of the original *B2A-S*, *X%B2A-S*, or *AFC* flow probability matrix, denoted by **V,** and the corresponding Null matrix, denoted by **N**:

$$\mathbf{C} = w\mathbf{V} + (1 - w)\mathbf{N}, \qquad 0 \le w \le 1 \tag{C-1}$$

where w is a weight placed on the original **V** matrix. Let passenger flow *probabilities* in cell *(i,j)* of the original matrix **V,** the corresponding null matrix **N**, and the corresponding combined matrix **C** be denoted by $p_{ij}^{(\mathbf{V})}$, $p_{ij}^{(\mathbf{N})}$, $p_{ij}^{(\mathbf{C})}$, respectively. For a given value of w, it follows from Eq. (C -1) that:

$$p_{ij}^{(\mathbf{C})} = w p_{ij}^{(\mathbf{V})} + (1 - w) p_{ij}^{(\mathbf{N})}, \quad 0 \le w \le 1 \tag{C-2}$$

Letting v_{ij} denote the passenger flow volume in the volume matrix from which the probability matrix **V** is derived, the multinomial probabilities, or likelihoods, of observing the v_{ij} values when using probabilities given in the **C** or **V** matrices are given in Eq. (C-3a) and Eq. (C-3b), respectively:

$$L(\mathbf{C}) = K \prod_{I=1}^{N} \prod_{J=1}^{N} p_{ij}^{(\mathbf{C})v_{ij}} \tag{C-3a}$$

$$L(\mathbf{V}) = K \prod_{I=1}^{N} \prod_{J=1}^{N} p_{ij}^{(V)v_{ij}} \tag{C-3a}$$

where K is a constant reflecting the combinatorics that is the same for the two likelihoods (since the v_{ij} volumes considered are the same), and the sums are taken over all cells with nonzero values. It can be shown that the probability or likelihood of observing the v_{ij} values is maximized when using Eq. (C-3b). The objective then is to form the **C** matrix, which replaces cells with probability values of 0 with nonzero probabilities, but not to reduce the likelihood of observing the "data" (either observed or sampled B2A

survey data or inferred B2A volumes inferred from the AFC data) "too much" as compared to the maximum likelihood given in Eq. (C-3b), that is, to make L(**C**) "close to" L(**V**).

Taking the ratio of the two likelihoods *LR* = L(**V**)/L(**C**) simplifies the calculations. From Eq. (C-3b) and Eq. (C-3a), this Likelihood Ratio is:

$$LR = L(\mathbf{V}) / L(\mathbf{C}) = \prod_{I=1}^{N} \prod_{J=1}^{N} (p_{ij}^{(\mathbf{V})} / p_{ij}^{(\mathbf{C})})^{v_{ij}} \qquad \text{(C-4)}$$

L(**C**) will be closer to L(**V**) as *LR* becomes closer to 1.(Since L(**V**) is the maximum likelihood, *LR* will be greater than 1, and L(**C**) will be closer to L(**V**) as *LR* becomes closer to 1 from above.) Multinomial likelihoods will be very small numbers, and L(**C**) will be orders of magnitude smaller than L(**V**). For this reason, and to allow a methodic determination of when L(**C**) is considered sufficiently close to L(**V**), the log-likelihood ratio *LLR* is formed as two times the (natural) logarithm of the likelihood ratio *LR*, which is equivalent to the (natural) logarithm of *LR* squared:

$$LLR = \ln(LR)^2 = 2\ln(LR) \qquad \text{(C-5)}$$

It follows that values of *LLR* closer to 0 will reflect that the **C** matrix is "closer to" the **V** matrix.

Decreasing the value of the weight *w* in Eq. (C-1) will lead to replacing the 0 probability values in the original **V** matrix with greater probability values. However, smaller values of *w* will make **C** and **V** more dissimilar, increasing *LLR*, which implies that it become less likely to observe the v_{ij} volumes with the new **C** matrix. That is, decreasing *w* will lead to a matrix **C** less in agreement with the original "data." As a result, the smallest value of *w* is sought such that *LLR* is greater than some threshold.

The threshold is determined in the spirit of Chi-Squared tests that are typically performed on log-likelihood ratios. (The term "in the spirit of" is used to emphasize that no claim is made that the *LLR* statistic satisfies the assumptions made when conducting statistical hypothesis tests on log-likelihood ratios.) Specifically, the threshold is taken as the Chi-Squared statistic value associated with degrees of freedom equal to the number of cells with probability equal to 0 in the original **V** matrix and significance level of 0.90. Although, there is no hypothesis testing in the choice of the *w* values, a significance level of 0.90 is selected as a commonly used value in hypothesis tests.

In typical Chi-Squared tests on the log-likelihood ratios, the number of degrees of freedom is equal to the difference in the number of restrictions on the two models leading to the likelihoods used in the ratio. The cells with probability flows of 0 in the original **V** matrix will all be assigned flow probabilities in the **C** matrix that depend on the value of *w* and the flow probabilities in the **N** matrix. The flow probabilities in the other cells will also depend on the possibly non-equal flow probabilities in the original **V** matrix. In this way, the number of cells in **V** matrix with probability equal to 0 can roughly be thought of as the difference in "restrictions" in the two matrices.

Therefore, the selection of *w* can be formulated as:

"Chose the minimum value of *w*, $0 \leq w \leq 1$, such that $LLR \leq$ *Chi-Square (df = # **V** zeros; α = 0.9).*"

Determining a better threshold than *Chi-Square (df = # **V** zeros; α = 0.9)* is certainly a topic for future research. Nevertheless, considering this threshold in the procedure is consistent with a few desirable properties. First, for the same **V** and **N** probability matrices and same value of *w*, the value of *LLR* will

increase as the total volume that leads to the **V** probability matrix increases. Therefore, a greater value of *w* will be required to make the **C** matrix more similar to the **V** matrix, thereby reducing the *LLR,* and satisfying the *LLR* ≤ *Chi-Square (df = # **V** zeros; α = 0.9)* constraint. Greater values of *w* lead to more weight being placed on the original **V** matrix, which is consistent with having more confidence in **V** matrix because of the greater amount of data used to determine the matrix.

Second, as the number of cells in the **V** matrix with flow probability of 0 increases, the *Chi-Square (df = # **V** zeros; α = 0.9)* threshold will increase. This will allow a lower value of *w* to be chosen (making the **V** and **C** matrices more dissimilar), thereby placing less weight on the **V** matrix. The motivation for developing a procedure is to replace cell values of 0 in the **V** matrix with "reasonable" values, implying that one does not have confidence in the 0 cell values. As such, it would follow that, all else equal, more cell values of 0 in the **V** matrix would lead to having less confidence in the **V** matrix.

Appendix D: SE&T Characteristics and their Definitions

Table D-1 lists all 34 SE&T characteristics considered in the expansion analysis of Section 5 along with their definitions of the categories of each.

Table D-1: SE&T characteristics used in the expansion analysis and their definitions

Description	Values	Values (continued)
Zip code where the respondent lives	Actual Value	
Type of place respondent is coming from now (O, origin)	1 = Your usual Workplace 2 = a Shopping place 3 = a School (K-12) 4 = a Hotel 5 = an Airport (as an air passenger) 6 = a Sporting event 7 = a Recreation / sightseeing place 8 = an Eating/Dining place 9 = a Medical appointment / doctor's visit	10 = a Social visit (friends/relatives) 11 = a College / University (students only) 12 =Your Home 13 = another business related place 14 = a place of Personal business (bank, post office) 15 = a place to Pick up/drop off someone (daycare, school) 16 = a convention / conference
Traffic analysis zone where the trip began	Actual Value	
Mode of access to transit	1 = Walked 2 = Biked 3 = Was dropped off by someone going someplace else 4 = Drove alone and parked	5 = Drove or rode with others and parked 6 = Wheelchair/scooter 9 = Other
Type of place respondent is going to now (D, destination)	1 = Your usual Workplace 2 = a Shopping place 3 = a School (K-12) 4 = a Hotel 5 = an Airport (as an air passenger) 6 = a Sporting event 7 = a Recreation / sightseeing place 8 = an Eating/Dining place 9 = a Medical appointment / doctor's visit	10 = a Social visit (friends/relatives) 11 = a College / University (students only) 12 = Your Home 13 = another business related place 14 = a place of Personal business (bank, post office) 15 = a place to Pick up/drop off someone (daycare, school) 16 = a convention / conference
Traffic analysis zone where the trip ended	Actual Value	

Table D-1, continued: SE&T characteristics used in the expansion analysis and their definitions

Description	Values	Values (continued)
Mode of egress from transit	1 = Walked 2 = Biked 3 = Was dropped off by someone going someplace else 4 = Drove alone and parked	5 = Drove or rode with others and parked 6 = Wheelchair/scooter 9 = Other
Traffic analysis zone where the respondent boarded the bus	Actual Value	
Number of transfers a respondent took before surveyed route from Origin and Number of transfers a respondent took after surveyed route to Destination	0 = None 1 = One 2 = Two 3 = Three or more	
Number of co-travelers on trip with respondent	0 = None (Zero) 1 = One (1) 2 = Two (2) 3 = Three (3) 4 = Four (4) 5 = Five (5)	6 = Six (6) 7 = Seven (7) 8 = Eight (8) 9 = Nine (9) 10 = Ten or more (10+)
Number of co-travelers on trip with respondent who are members of respondents household	0 = None (Zero) 1 = One (1) 2 = Two (2) 3 = Three (3) 4 = Four (4) 5 = Five (5)	6 = Six (6) 7 = Seven (7) 8 = Eight (8) 9 = Nine (9) 10 = Ten or more (10+)
Did respondent/will respondent make this trip in exactly the opposite direction today	1 = Yes 2 = No	
Average days per week using public transit	0 = None / Never 1 = One day a week 2 = Two days a week 3 = Three days a week	4 = Four days a week 5 = Five days a week 6 = Six days a week 7 = Seven days a week
Payment method of respondent	1 = Cash Fare 2 = All Day Pass 3 = 1-trip ticket 4 = 2-trip ticket 6 = 10-trip ticket 7 = 7-Day Pass 8 = 31-Day Pass	9 = Columbus Public/Metro School Student ID 10 = OSU/CCAD College ID 11 = COTA/County Employee 12 = Guest Pass 13 = Other 14 = Not provided
Fare discounts received	1 = None 2 = Seniors/Key/ADA 3 = Student (9-12) 4 = Student (College)	5 = Child over 48 inches tall, under 12 years old 6 = Child under 48 inches tall accompanied by adult 9 = Other

Table D-1, continued: SE&T characteristics used in the expansion analysis and their definitions

Description	Values	Values (continued)
Respondent lives in Central Ohio	1 = Yes 2 = No	
Number of working vehicles available to respondent household	0 = None (0) 1 = One (1) 2 = Two (2)	3 = Three (3) 4 = Four or more (4+)
Number of household members	1 = One (1) 2 = Two (2) 3 = Three (3) 4 = Four (4) 5 = Five (5)	6 = Six (6) 7 = Seven (7) 8 = Eight (8) 9 = Nine (9) 10= Ten or More (10+)
Number of employed household members	1 = One (1) 2 = Two (2) 3 = Three (3) 4 = Four (4) 5 = Five (5)	6 = Six (6) 7 = Seven (7) 8 = Eight (8) 9 = Nine (9) 10 = Ten or More (10+)
Respondent employment status	1 = Employed full-time 2 = Employed part-time 3 = Not currently employed but seeking work	4 = Not currently employed and not seeking work 5 = Retired 6 = Homemaker
Respondent student status	1 = Not a student 2 = Yes, Full Time college/university 3 = Yes, Student through 12th grade	4= Yes, Part Time college/university 5 = Yes, Other
Does respondent have a valid driver's license	1 = Yes 2 = No	
Does respondent have a disability	1 = Yes, ADA Certified disability 2 = Yes, other disability 3 = No	
Respondent age	1 = Under 16 2 = 16 to 17 3 = 18 to 24 4 = 25 to 34	5 = 35 to 49 6 = 50 to 64 7 = 65 to 74 8 = 75 or Older
Respondent indicated whether or not they are Asian	Yes or No	
Respondent indicated whether or not they are Native American	Yes or No	
Respondent indicated whether or not they are Black	Yes or No	
Respondent indicated whether or not they are Hispanic	Yes or No	
Respondent indicated whether or not they are Somali	Yes or No	

Table D-1, continued: SE&T characteristics used in the expansion analysis and their definitions

Description	Values	Values (continued)
Respondent indicated whether or not they are White	Yes or No	
Gender of respondent	1 = Male 2 = Female	
Total annual household income in 2012 before taxes	1 = Less than $10,000 2 = $10,000 to $14,999 3 = $15,000 to $24,999 4 = $25,000 to $49,999	5 = $50,000 to $74,999 6 = $75,000 or More 7 = Don't Know/Refused
Determined using origin place type and destination place type	Homebased work = O: 4, 12 to D: 1, 13 and reverse Homebased school = O: 4, 12 to D: 3 and reverse Homebased College/University = O: 4, 12 to D: 11 and reverse Homebased errands = O: 4, 12 to D: 2, 9, 14, 15 and reverse Homebased entertainment = O: 4, 12 to D: 6, 7, 8, 10 and reverse Homebased to hotel = O: 4, 12 to D: 4, 12 and reverse Homebased other = O: 4, 12 to D: 5 and reverse Work to errand = O: 1, 13 to D: 1, 2, 3, 5, 6, 7, 8, 9, 10, 13 and reverse Nonhome to nonhome = O: 1, 2, 3, 5, 6, 7, 8, 9, 10, 13 to D: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and reverse	

Appendix E: Expansion Process and Example

In order to take into account the true B2A flows of the route, and correct for any under- or over-sampling of certain B2A pairs in the OBS survey, each characteristic is expanded using the expansion B2A flow probability matrix. This calculation involves creating a matrix for each category of a characteristic showing the probability of that characteristic taking the value of that category. Each probability value is then multiplied by the true expansion B2A probability values to calculate the expected proportion of riders who traveled each B2A pair with each value of the characteristic. The values are then added together by category, providing the expected proportion for each category of each characteristic across the population traveling on that route-direction and period. This is done for every characteristic and yields the expanded proportions and flows for each route-direction and period.

What follows is an example that illustrates this process:

Using the hypothetical values shown and calculated in the example of Appendix F (for synthetic record creation), this example demonstrates how each characteristic is expanded. Table E-1 is the expansion probability B2A matrix, for a certain route-direction and period.

Table E-1: Expansion B2A probability matrix

		Alighting			
Boarding		1	2	3	Total Ons
	1	0.130	0.058	0.174	0.362
	2		0.174	0.355	0.529
	3			0.109	0.109
Total Offs		0.130	0.232	0.638	

The relative proportion of each disability category by cell, as determined by the OBS survey, is shown on the left side of Table E-2 below. Note that this includes the synthetic values calculated in the example of Appendix F. Each cell is expanded by multiplying the value by the expansion B2A probability of that cell shown in Table E-1. The result of this expansion is shown on the right side of Table E-2. The value in each cell represents the proportion of all passengers (in that route-direction and period) who had that disability value and B2A pair.

The cells of each matrix on the right side of Table E-2 can be added to produce the expected proportion for each category of each characteristic. These results are shown in Table E-3, alongside the proportions taken from the OBS survey prior to expansion. These proportions can then be multiplied by total ridership for that route-direction and period, and each characteristic is expanded using this method.

Table E-2: Relative category proportions and expected proportions after expansion

Disability Category Proportions				
Disability = 1				
		Alighting		
Boarding		1	2	3
	1	0.500	0.442	0.385
	2		0.444	0.200
	3			0.286
Disability = 2				
		Alighting		
Boarding		1	2	3
	1	0.125	0.198	0.308
	2		0.111	0.267
	3			0.429
Disability = 3				
		Alighting		
Boarding		1	2	3
	1	0.375	0.359	0.308
	2		0.444	0.533
	3			0.286

Expected Proportion (“Expanded”)				
Disability = 1				
		Alighting		
Boarding		1	2	3
	1	0.065	0.026	0.067
	2		0.077	0.071
	3			0.031
Disability = 2				
		Alighting		
Boarding		1	2	3
	1	0.016	0.011	0.054
	2		0.019	0.095
	3			0.047
Disability = 3				
		Alighting		
Boarding		1	2	3
	1	0.049	0.021	0.054
	2		0.077	0.189
	3			0.031

Table E-3: Expected Proportions after Expansion Compared to Main Survey original Proportions

Category	Expected Proportion (after Expansion)	Proportions from Main Survey (before Expansion)
Disability = 1	0.337	0.346
Disability = 2	0.242	0.250
Disability = 3	0.421	0.404

Appendix F: Explanation and Example of Synthetic Record Creation in the OBS B2A matrices

Synthetic records in the OBS B2A matrices need to be created when the B2A matrix to be used for expansion indicates a certain B2A pair was traveled but the OBS survey does not include SE&T characteristic data for that B2A pair. The synthetic records are created for each characteristic using the boarding and alighting marginal totals corresponding to the B2A pair where a synthetic record needs to be created. First, the information for each characteristic is divided by driver's license category given that this characteristic fundamentally distinguishes choice and captive transit riders. After being split by driver's license, the proportion of each category in a characteristic is determined. These proportions are then weighted by the overall driver's license proportion from the OBS survey. This weighted proportion is the final synthetic record value for the characteristic.

The following example demonstrates the basic steps of this process, which is repeated for every characteristic:

Step 1: Identify a need for creating a synthetic Record

A synthetic record needs to be created when the B2A matrix for a certain route-direction and period to be used for expansion indicates a certain B2A pair was traveled, but the OBS survey does not include characteristic information from any passengers. Table F-1 shows an example of when a synthetic record needs to be created, with the highlighted cell indicating a cell where no characteristic data were collected. This cell, which represents passengers who boarded at segment 1 and alighted at segment 2, now shows "NEED" to make clear that a synthetic record is necessary.

Table F-1: B2A/APC O-D Matrix and Main Survey O-D Matrix showing Synthetic Record Needed

OBS survey B2A matrix

		Alighting			
Boarding		1	2	3	Total Ons
	1	8	NEED	13	21
	2		9	15	24
	3			7	7
Total Offs		8	9	35	

Expansion B2A matrix

		Alighting			
Boarding		1	2	3	Total Ons
	1	18	5	24	47
	2		24	34	58
	3			15	15
Total Offs		18	29	73	

Table F-2 shows hypothetical data for the Disability characteristic, which has three categories. The OBS survey produced this sort of data for each category of each characteristic, and synthetic records for all characteristics were calculated using the method illustrated in this example.

Step 2: Identify Marginal Boardings (Ons) and Alightings (Offs) associated with each synthetic record needed and split by Driver's License equal to Yes (1) or No (2)

Each characteristic and category is split by the records that have driver's license equal to yes (1) or no (2). The marginal boarding (segment 1) and marginal alighting (segment 2) information that will be used to create the synthetic record is identified (yellow and orange cells below). The total proportion of passengers who had driver's licenses and who didn't have driver's licenses was also determined, as shown by the DL1 Factor and DL2 Factor values at the bottom of Table F-3.

Table F-2: Data collected for each category of the Disability

OD Matrix for Disability = 1					
		Alighting			
Boarding		1	2	3	Total Ons
	1	4	NEED	5	9
	2		4	3	7
	3			2	2
Total Offs		4	4	10	

OD Matrix for Disability = 2					
		Alighting			
Boarding		1	2	3	Total Ons
	1	1	NEED	4	5
	2		1	4	5
	3			3	3
Total Offs		1	1	11	

OD Matrix for Disability = 3					
		Alighting			
Boarding		1	2	3	Total Ons
	1	3	NEED	4	7
	2		4	8	12
	3			2	2
Total Offs		3	4	14	

Step 3: Calculate Proportion of each category and Proportion weighted by Driver's License

As shown in Table F-4, the marginal boarding and alighting numbers identified in Table F-3 are added together to produce a total associated with each cell where a synthetic record is needed. The proportion of each category, still split between driver's license equal to yes or no, is then determined. The weighted proportion for each category is then calculated by multiplying each proportion shown in Table F-4 by the associated "DL Factor" found in Table F-3, to produce the values below in Table F-5. These values are the final synthetic values.

Table F-3: Disability Categories split by Driver's License with Marginal Ons and Offs identified

Disability with Driver's License = 1

OD Matrix for Disability = 1					
		Alighting			
Boarding		1	2	3	Total Ons
	1	2	NEED	1	3
	2		2	0	2
	3			1	1
Total Offs		2	2	2	

OD Matrix for Disability = 2					
		Alighting			
Boarding		1	2	3	Total Ons
	1	1	NEED	2	3
	2		0	1	1
	3			2	2
Total Offs		1	0	5	

OD Matrix for Disability = 3					
		Alighting			
Boarding		1	2	3	Total Ons
	1	1	NEED	2	3
	2		3	3	6
	3			0	0
Total Offs		1	3	5	

Total DL = 1	21	
DL1 Factor	0.404	

Disability with Driver's License = 2

OD Matrix for Disability = 1					
		Alighting			
Boarding		1	2	3	Total Ons
	1	2	NEED	4	6
	2		2	3	5
	3			1	1
Total Offs		2	2	8	

OD Matrix for Disability = 2					
		Alighting			
Boarding		1	2	3	Total Ons
	1	0	NEED	2	2
	2		1	3	4
	3			1	1
Total Offs		0	1	6	

OD Matrix for Disability = 3					
		Alighting			
Boarding		1	2	3	Total Ons
	1	2	NEED	2	4
	2		1	5	6
	3			2	2
Total Offs		2	1	9	

Total DL = 2	31	
DL2 Factor	0.596	

Table F-4: Calculating Proportion for each "NEED" cell

		By Boarding	By Alighting	Total	Proportion
Driver's License = 1	Disability = 1	3	2	5	0.357
	Disability = 2	3	0	3	0.214
	Disability = 3	3	3	6	0.429
Driver's License = 2	Disability = 1	6	2	8	0.5
	Disability = 2	2	1	3	0.1875
	Disability = 3	4	1	5	0.3125

Table F-5: Proportion of each Disability Category after weighted by Driver's License

Category	Weighted Proportion
Disability = 1	0.442
Disability = 2	0.198
Disability = 3	0.359