Investigating Sampling Biases in Transit Onboard Surveys and Associated Impacts on Resulting Passenger Socioeconomic and Travel Characteristics

Rabi G. Mishalani · Mark R. McCord · Andre Carrel · Jakob A. zumFelde

Abstract Conducting transit onboard surveys is a primary way that transit agencies and metropolitan planning organizations (MPOs) gather data on transit user socioeconomic and travel (SE&T) characteristics. These surveys are known to suffer from a variety of possible sampling and response biases. This study investigates two possible causes of sampling biases: those resulting from passenger making short trips who, as a result, do not have sufficient time to complete an interview or questionnaire, and those resulting from a mismatch between the number of surveyors and the passenger boarding volumes. A large-scale empirical study was conducted to investigate the presence of these biases. The results indicate that the short-trip bias remains present despite efforts to mitigate it while the bias associated with surveyor workload is mitigated. Furthermore, the presence of short-trip bias appears to vary by route and time-of-day. Finally, short trips are found to be associated with travelers who are making shopping trips, of low income, who are...
non-white, or who are transit dependent, whereby such travelers are more likely to be under-represented by onboard surveys.

**Keywords:** Transit Passenger Socioeconomic and Travel Characteristics · Transit Onboard Surveys · Short-trip Sampling Bias · Surveyor Workload Sampling Bias

1 Introduction

Conducting transit onboard surveys is a primary way that transit agencies and metropolitan planning organizations (MPOs) gather data on transit user socioeconomic and travel (SE&T) characteristics. The collected data are used for a variety of reporting and planning purposes. These surveys are known to suffer from a variety of possible sampling and response biases.

The objective of this study is to test for the presence of two possible causes of sampling biases: those resulting from passenger making short trips who, as a result, do not have sufficient time to complete an interview or questionnaire while onboard a transit vehicle leading to under-sampling, and those resulting from a mismatch between the number of surveyors and the passenger boarding volumes leading to under- or over-sampling. In the case of the presence of short-trip bias, additional objectives include the investigation of whether this bias is associated with certain route and time-of-day periods and whether short trips are associated with certain traveler socioeconomic and travel characteristics.

2 Data

Three types of data on several Central Ohio Transit Authority (COTA) routes were collected and analyzed as part of this study. In what follows the various datasets 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. The OBS data represent 7,987 passengers and 12,512 passenger trips (many interviewed passengers conducted round-trips). This sample size is approximately equal to the target of 10% of the average daily ridership for each route-direction and time-of-day period. In collecting these data, concerted efforts were made to mitigate any sampling bias resulting from traveler making short trips. Specifically, when a randomly selected passenger agrees to participate upon boarding a bus, the interviewer asked if the respondent had at least five minutes to complete the survey. If the
passenger responded negatively, the interviewer asked for the person’s home address, boarding and alighting location, name, and phone number. Within three days of the trip, a phone interviewer then called the person to ask the survey questions. If the person had at least five minutes on the bus, the interviewer administered the questionnaire in the form of a face-to-face interview. If the passenger was unable to answer all of the questions prior to alighting the bus, their phone number was recorded and they were called afterwards to answer the remaining survey questions. Moreover, the design of the survey procedure was such that appropriate numbers of surveyors were allocated to certain portions of routes during certain time-of-day periods on the basis of the expected number of boarding passengers. Specifically, the allocation was intended to reduce the chance of having too many surveyors and, thus, over-sampling passengers at certain boarding stops and time-of-day periods or having too few surveyors and, thus, under-sampling passengers at certain boarding stops and time-of-day periods.

**Boarding-to-alighting OD passenger flow survey data (OD-S):** Stop-to-stop OD 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. Sample sizes that mostly exceed the average daily ridership for each route-direction and time-of-day are employed. Such a sample sizes is considered large with respect to the 20% of the average daily ridership commonly employed in practice. The OD-S data captured the boarding-to-alighting stop pairs of 45,324 passengers traveling on 2,420 bus trips.

**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 – are used in this analysis. 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 various datasets required multiple types of processing to address measurement errors and other types of problems identified through extensive data quality investigations.
3 Methodology and Analysis

The data are used to determine several variables for the purpose of the subsequent analysis. The APC and OD-S data are combined to estimate OD flow matrices for each route-direction and time-of-day period. Stops are grouped into pre-defined route segments and the estimated OD volumes are normalized to produce probability OD matrices where a cell entry represents the probability that a random passenger traveling on the bus route-direction during the time-of-day period would travel from the boarding segment to the alighting segment indicated by the OD cell. The resulting OD flow matrices are considered to represent the “ground truth” in this analysis.

The OBS data are used to determine the SE&T characteristic for each of the OD pair cells described above. In addition, the normalized (probability) OD flows for the respondents to the SE&T survey are determined for the same route-directions, time-of-day periods, and OD cells based on the same stop segments.

The SE&T OD flows and the estimated “true” OD flows are compared to determine whether each OD cell is over- or under-sampled. The difference between the two probability OD flows for each cell is used to capture this comparison. A negative difference between an SE&T probability OD flow and the estimated “true” probability OD flow reflects under-sampling and a positive difference reflects over-sampling. These differences for each OD cell are converted to a binary indicator variable where negative differences are mapped into an indicator of 1 and positive differences into an indicator of 0 representing the presence of under- and over-sampling for each OD cell, respectively.

An important aspect of this analysis is the definition of which OD pairs should be considered reflective of “short trips” given the focus on short-trip bias. In this study, short trips are defined as trips that began and ended in the same route segment. A binary indicator variable is defined to take the value of 1 if an OD cell is considered to represent a short trip based on this definition (i.e., a cell that belongs to the diagonal of the OD flow matrix), and 0 otherwise (i.e., a cell that belongs to the off-diagonal of the OD flow matrix).

Given the interest in the effect of surveyor workload on over- and under-sampling, another important variable is the hourly volume of passengers traveling on each OD pair, which are calculated from the “true” OD flows
estimated from the APC and OB-S data prior to the normalization into probability OD flows and the duration of the corresponding time-of-day period.

The remaining variables that are considered are categories of the characteristics that were asked about in the SE&T survey. The survey collected data on more than 30 SE&T characteristics, with each having at least two categories. For each OD pair, the value used for the variable capturing whether a characteristic is over or under represented in an OD cell is the difference between the percentage of passengers in a category for each OD pair and the percentage of passengers of that category for the route-direction and time-of-day period. Defining such a relative measure is important given that the data used in the analysis are pooled across multiple route-directions and periods.

Three types of binary logit models are estimated based on the above variables. To investigate the presence of short-trip and surveyor workload sampling biases, the first type of model considers the over- and under-sampling indicator variable as the dependent variable and the short-trip indicator and passenger hourly volume variables as explanatory variables. To investigate whether short-trip biases are associated with route and time-of-day characteristics, the second type of model also considers the over- and under-sampling indicator variable as the dependent variable but only for short trip OD pairs. The explanatory variables in this case are route and time-of-day period indicators. To investigate the possible association of short-trips with certain SE&T characteristic, the third type of model considers the short-trip indicator as the dependent variable and the relative representation of each SE&T characteristic in each OD cell. Various variations of each type of model are estimated and interpreted based on the signs of the estimated coefficients and the statistical significance of the coefficients.

4 Results and Conclusions

The empirical results indicate that the short-trip bias remains present despite efforts to mitigate it while the bias associated with surveyor workload is effectively mitigated by the appropriate allocation of surveyors to route-directions and time-of-day periods. Furthermore, the presence of the short-trip bias appears to vary by route and time-of-day. Finally, short trips are found to be associated with travelers who make shopping trips, of low income, who are non-white, or who are transit dependent. More specifically,
such travelers are the more likely to be under-represented by onboard surveys due to the short-trip sampling bias.