Innovative Sampling Plans for Estimating Transit Passenger-Kilometers

PETER G. FURTH

Estimating passenger-kilometers or passenger-miles to meet National Transit Database requirements usually involves costly sampling. Three innovative sampling plans are described that have been developed to reduce sampling requirements. The first method, which proved to be very effective when total boardings is known, uses a small number of ride checks (ons and offs by stop) on each route. Average trip length is estimated as a combined ratio estimator from a stratified sample. The second method was applied where the boardings total is not known. It uses both a sample of ride checks and another sample (needed for another purpose) that measures only boardings. A "mixed estimator" is derived that optimally combines two separate estimators: a simple mean from the ride check sample and an average trip length from the ride check sample multiplied by average boardings from the other sample. This second method proved effective for a single light-rail line but only marginally effective for a large bus system with widely varying route lengths. The third method exploits the pattern of symmetry in boarding and alighting patterns in opposite directions to estimate average trip length by route using boardings data only. Average trip length is the algebraic difference between the boardings centroids in the two directions. For the two routes analyzed, this method turned out to be ineffective in comparison with other methods because of high between-trip variability in the boardings centroids.

To comply with FTA guidelines for the National Transit Database (NTD), transit agencies are required to report annual passenger boardings and passenger-kilometers (or passenger-miles) by mode. Some agencies know total boardings because they count every boarding passenger, but many agencies do not. And most transit agencies do not know passenger-kilometers because they do not record each passenger's on stop and off stop. Agencies that do not know passenger boardings and passenger-kilometers from routine counts must estimate these quantities from a sample. The estimates are required to attain a specified level of accuracy: a precision of ±10 percent at the 95 percent confidence level. The sampling process is generally manual and expensive, involving on-board surveyors called "checkers," who perform "ride checks," recording the number of passengers getting on and off at each stop for a set of sampled trips.

AVAILABLE SAMPLING PLANS

At this time only one default sampling plan has been approved by FTA for general use on bus systems; it is described in FTA Circular 2710.1A (*1*). Agencies may use other sampling plans if they have a statistician certify that they meet the specified accuracy criteria. The default plan, which includes a few alternatives, calls for the random sampling of at least 549 single bus trips. On each selected trip, a ride check is conducted, from which total boardings and passenger-kilometers for the trip are calculated. This sampling plan is based on simple expansion of the sample means. Necessary sample sizes are based on default estimates of trip-level coefficients of variation of passenger-kilometers, which are typically more variable than boardings.

The expense involved in this sampling process can be considerable. For example, sampling a single 30-min trip usually requires far more than 30 min. Checkers using a car nearly always have to make a round trip in order to return to their vehicle; checkers without a car need time to ride to the start of the selected trip and then to their next duty after checking the selected trip. It is not unusual, then, that 0.5 or more full-time equivalent employees are used for NTD ride checks.

In an effort to improve sampling efficiency, a revenue-based sampling plan was published by FTA in 1985 (*2*) that required only 208 trips a year, provided that cash revenue could be recorded for each sampled trip along with ons and offs by stop. This method involves estimating from the sample the ratio of passenger-kilometers to cash revenue and expanding by annual total cash revenue. The smaller sample size was justified by the strong correlation between cash revenue and passenger-kilometers. However, with the widespread adoption of passes, tickets, and other forms of prepayment, cash revenue has grown to be a less reliable indicator of trip patronage, and thus FTA no longer approves this plan by default, although it may still be used if certified by a statistician.

Other sampling plans have been developed by various transit agencies, either to reduce their sampling cost or to satisfy more stringent accuracy criteria. An early example developed by Phifer (*3*) is based on a regression estimator. A data collection manual published by FTA (*4*) encourages statistical estimation of route-level measures for improved management and planning. Furth and McCollom (*5*) describe the application of ratio estimators for improving sampling efficiency. Furth et al. (*6*) discuss the benefits of sampling by a cluster of trips (e.g., round trips or a 4-h chain of trips on a single route), a technique that improves efficiency by reducing the overhead associated with sampling each trip. The widespread use of electronic fareboxes, which, in some systems, provide reliable boardings counts, has led to sampling plans that take advantage of this information. For example, Huang and Smith (*7*) explored various cluster sizes for NTD sampling in the presence of complete boardings counts and found that an efficient sampling plan involved round trips and a ratio of passenger-kilometers to boardings. Furth and Kumar (*8*) describe the application of two-stage sampling in the context of a single light-rail line (without a farebox) requiring accurate patronage estimates.

It is convenient at this point to state the sample size formula for single-stage, single-stratum sampling:

$$n = \frac{t^2(cv)^2}{(prec)^2} \quad (1)$$

Northeastern University, Department of Civil and Environmental Engineering, 360 Huntington Ave., Boston, MA 02115.

where n is the sample size required to achieve a specified precision (expressed as a decimal, e.g., $prec = 0.1$ for ±10 percent precision) at a confidence level for an associated t-value (e.g., for the 95 percent confidence level, $t = 1.96$ when the sample size is large, and it rises above 2.04 when the sample size is smaller than 30). Besides these two parameters, which depend on the specified accuracy level, necessary sample size also depends on coefficient of variation (cv) of the variable being estimated. This formula can be applied with simple expansion of the sample mean, in which case cv is the cv of passenger-kilometers. It can also be applied with ratio estimation using the unit cv of the ratio, as described by Furth and McCollom (*5*).

MOTIVATION FOR NEW SAMPLING PLANS

In the last two years, FTA and the American Public Transit Association have sponsored a program called the Transit Passenger Monitoring System (TPMS) that encourages transit agencies to implement a regular program of surveying passengers using a short self-service questionnaire to help determine what benefits the passengers are getting from using transit. Part of the rationale for TPMS is that it should follow a statistically valid sampling plan, with sampling spread over the whole year. It made sense, therefore, to coordinate sampling for TPMS with NTD sampling. Because TPMS was implemented in several cities, there was the opportunity to develop improved NTD sampling plans as well as TPMS sampling plans and to coordinate sampling for the two programs. In this paper three innovative sampling plans are described that were developed in the course of this project.

The first sampling plan deals with a small transit system of eight routes that routinely counts boardings. Because of the small number of routes in the system, it is easy to ensure that the ride check sample covers all of the routes and thus to treat it as a stratified sample, eliminating most of the between-route variability. A technique called *combined ratio estimation* was applied because it permits small sample sizes per route, resulting in a very efficient sampling plan. This technique should have wide applicability to systems of up to 50 routes. The second sampling plan involves a large city that needs to estimate both boardings and passenger-kilometers by sampling. A new sampling technique was developed that combines data from ride checks with boardings counts that are made in the course of TPMS sampling. This method proved effective for a single light-rail line but not for a large and varied bus system. The third sampling plan involves a new method to estimate average trip length on a light-rail line by recording ons by stop only, on the basis of the concept of symmetry in boarding and alighting patterns in opposite directions of a route. This method turned out to be comparatively ineffective for the two routes analyzed, although it could have application in other contexts where boarding patterns are less variable.

ROUTE-LEVEL STRATIFICATION WITH COMBINED RATIO ESTIMATION

Kenosha (Wisconsin) Rapid Transit (KRT) is a small, eight-route system. Bus operators count passengers on every trip. Passenger-kilometers is estimated from a sample of ride checks, from which average (passenger) trip length (ATL), the ratio of passenger-kilometers to boardings, is estimated and then expanded by annual boardings. With only eight routes, it is clear that a ride check sample of any reasonable size can easily cover all of the routes. Because it is typical of transit systems that most of the variation in average trip length is between rather than within routes, stratification by route can eliminate some or all of the effect of between-route variation, thereby reducing the needed sample size.

Stratified Sampling and Combined Ratio Estimation

One approach to stratifying by route, which at the same time takes advantage of the available data on boardings by route, is to estimate ATL for each route, expand it by route boardings, and aggregate over all routes. On the surface this method, called *stratified ratio estimation,* appears to be a very efficient method, because it eliminates all the between-route variation. However, the effectiveness of this method is limited by the need to avoid bias. The bias associated with ratio estimators does not become negligible until the sample is at least of moderate size (*9*). One analysis of transit ridership data resulted in the recommendation of a minimum of 10 samples for ratio estimates to avoid significant biases (*4*). Because this method involves estimating and expanding a ratio for each route, a lower limit for the sample size is 10 samples per route, or 80 trips overall for KRT. Analysis of KRT data showed that this number of trips is more than needed to meet NTD accuracy requirements and represents a large savings compared with 550 trips in the default plan.

In fact, if bias in the ratio estimates could be ignored, the necessary sample size would be only two or three trips per route. This result led to the exploration of a related stratified sampling technique known as *combined ratio estimation* (*9*). It is not quite as efficient as stratified ratio estimation, because it does not eliminate all of the between-route variation, but because it is not as subject to bias, it can permit smaller sample sizes. The combined ratio estimator is found using the following steps:

1. For each stratum (route), find the sample mean passenger-kilometers and boardings;
2. Find the estimated stratum total passenger-kilometers and boardings by expanding the stratum sample means by the total number of trips in the stratum;
3. Find the estimated passenger-kilometers and boardings grand totals by summing the estimated stratum totals;
4. Take the combined ratio, which is the ratio of the estimated passenger-kilometers grand total to the estimated boardings grand total; and
5. Expand the combined ratio by total system boardings to yield estimated total passenger-kilometers.

Because each of the estimated stratum totals is unbiased (regardless of sample size), and because the number of samples involved in calculating the ratio when it is finally taken in Step 4 is much larger than the number that would be involved in a single stratum's ratio, the bias associated with the combined estimator may be considered negligible when the overall sample size is more than 30. Another advantage of this technique is that it can be applied even if a transit system has only system-level, not route-level, boardings data.

Variance of the Combined Ratio Estimator

The variance of the combined ratio estimator depends in part on how the sample is allocated between strata. Although it is possible to

determine an optimal allocation, it is usually the case with stratified sampling that proportional allocation (distributing the sample among the strata in proportion to stratum size, i.e., the number of 2-h cycles in each route's daily schedule) is nearly as efficient as optimal allocation, and it is simpler for sample selection, expansion, and analysis. With proportional allocation, the unit *cv* of the combined ratio estimate (defined as the square root of the relative variance per sample) can be estimated from historical data by

$$cv = \frac{1}{\bar{y}}\sqrt{\sum_h w_h\left(s_{yh}^2 + R^2 s_{xh}^2 - 2Rr_{xyh}s_{yh}s_{xh}\right)} \qquad (2)$$

where

h = stratum,
$\bar{y}$ = passenger-kilometers,
x = boardings,
s = sample standard deviation,
r = sample correlation coefficient,
w_h = relative (population) size of stratum h (such that the sum of the w_h's is 1), and
R = combined ratio = $\bar{y}/\bar{x}$, where the grand means, using the standard estimate based on stratified sampling, are

$$\bar{y} = \sum_h w_h y_h \qquad \bar{x} = \sum_h w_h x_h$$

Results and Discussion

An analysis of ride check data from four of KRT's routes revealed that the within-route variation in average trip length is indeed very small. The data available were one weekday ride check for every scheduled trip in the system. To make sampling more cost-effective, the sampling unit chosen was not the single trip, but a "2-h cycle," which on most routes is simply a round trip and on the shorter routes is a chain of two round trips. The data were therefore aggregated by 2-h cycle, with most routes having 16 such cycles in a weekday. The within-route levels of variation, as measured by the unit *cv* of the passenger-kilometers to boardings ratio, are given in Table 1. As indicated in the data, route-level unit *cv*'s of average trip length were between 0.10 and 0.26. For comparison, Huang and Smith's work (*7*) in Madison, Wisconsin, found that without stratification by route, the unit *cv* of average trip length at the round trip level was 0.36, and experience with other transit systems suggests that values as high as 0.60 are not unusual.

On the basis of an analysis of four of the eight KRT routes, the unit *cv* for the combined estimator was found to be 0.19. With this low a unit *cv*, the necessary sample size to achieve the NTD accuracy standard is, following Equation 1, only 18 two-hour cycles a year. To provide a margin of safety, the sampling plan recommended in this study calls for one 2-h cycle per week, or 52 per year. This represents a dramatic savings compared with the 550 or so trips that are called for by Circular 2710.1A and a moderate savings over the stratified ratio estimation method, which required 80 samples per year. The resulting precision, at the 95 percent confidence level, is ±5.4 percent.

There is good reason to expect this method to be similarly effective in other transit systems with up to 50 routes, a small enough size that the ride check sample can provide at least three round trips on each route. Beyond 50 routes, other methods that do not require sampling on every route are likely to be more efficient.

MIXED ESTIMATOR USING PAIRED SAMPLE AND BOARDINGS-ONLY SAMPLE

The Niagara Frontier Transportation Authority (NFTA) in metropolitan Buffalo is representative of a transit agency that does not have daily route-level passenger counts on either its bus system or its light-rail line. Revenue-based estimation is not possible either, for practical reasons. Therefore, NFTA practice has been to estimate both boardings and passenger-kilometers by simple expansion of the mean from a sample of ride checks. On the basis of historical data from single-trip samples, simple expansion of sample means requires a minimum of 450 single-trip ride checks for bus and 197 for rail to achieve NTD accuracy.

The TPMS project presented an opportunity for reducing the NTD ride check sample size by taking advantage of boardings counts (used as control totals) obtained from the TPMS samples. These boardings counts offer additional information about the mean boardings per trip, which when combined with an average trip length ratio estimated from ride checks provides a better estimate of passenger-kilometers than could be obtained from the ride checks alone. An estimation method called the *mixed estimator* was developed to make optimal

TABLE 1 Route-Level Statistics and Combined Ratio Estimator Results for Kenosha Rapid Transit

Route	1	2	5	6	Combined
size	16	16	16	8	
relative size	0.286	0.286	0.286	0.143	**1**
mean boardings	33	51	53	36	
boardings	0.542	0.440	0.417	0.646	
cv pass-km	0.633	0.526	0.469	0.509	
correlation	94.1%	97.4%	98.2%	92.2%	
unit cv of ratio	0.221	0.139	0.098	0.264	**0.193**
necessary n					**18**
recommended n	15	15	15	7	**52**
precision					**5.4%**

use of the available information. Its derivation, given in the following two subsections, can be skipped without loss of continuity.

Derivation of Mixed Estimator

A formal presentation of the sampling and estimation method follows. Let

X_i, Y_i = boarding (passenger-km) on trip i

The quantity to estimate is

$\bar{Y}$ = population mean passenger-kilometers per trip

There is a sample of n_{I} ride checks, each consisting of a paired observation of X_i and Y_i, with sample means

$\bar{x}, \bar{y}$ = sample mean of $\bar{X}, \bar{Y}$ from the ride check sample

from which an estimate of the average trip length is obtained,

$$R_{\text{I}} = \frac{\bar{y}_{\text{I}}}{\bar{x}_{\text{I}}}$$

There is also a TPMS sample of n_{II} trips for which only boardings is measured:

$\bar{x}_{\text{II}}$ = sample mean of X from the TPMS sample

From these basic statistics, one can obtain two estimators of the mean passenger-kilometers per trip:

$$\hat{y}_{\text{I}} = \bar{y}_{\text{I}} \qquad \hat{y}_{\text{II}} = \bar{x}_{\text{II}} R_{\text{I}} = \bar{x}_{\text{II}} \frac{\bar{y}_{\text{I}}}{\bar{x}_{\text{I}}} \tag{3}$$

The first estimator is obviously unbiased. The second may also be considered unbiased, because the bias associated with ratio estimates is negligible provided the sample size is large (*9*), which is the case in this application. A general estimator, called the *mixed estimator,* is a weighted sum of the previous two estimators:

$$\hat{y}_3 = (1 - w)\hat{y}_1 + w\hat{y}_2 \tag{4}$$

In this formula, w can be any constant between 0 and 1 and may be selected to minimize the variance of the estimator.

Variance of Mixed Estimator

The variance of the mixed estimator is the sum of three terms:

$$V(\hat{y}_3) = (1 - w)^2 V(\hat{y}_1) + w^2 V(\hat{y}_2) + 2w(1 - w)\ \text{Cov}(\hat{y}_1, \hat{y}_2) \tag{5}$$

To further develop Equation 5, note first that, from simple random sampling,

$$V(\hat{y}_1) = \frac{v_y^2 \bar{Y}^2}{n_{\text{I}}}$$

in which v stands for the coefficient of variation. For the second term, assuming independence between the ride check and the boardings count samples,

$$V(\hat{y}_2) = V(R_{\text{I}}\bar{x}_2) = E^2(R_{\text{I}})V(\bar{x}_2) + V(R_{\text{I}})E^2(\bar{x}_2)$$

$$= \frac{R^2 X^2 v_x^2}{n_{\text{II}}} + \frac{R^2 X^2 u_R^2}{n_{\text{I}}} = \bar{Y}^2\left(\frac{v_x^2}{n_{\text{II}}} + \frac{u_R^2}{n_{\text{I}}}\right)$$

where u_R is the unit (i.e., per sampled trip) coefficient of variation of the average trip length ratio, given by (*9*)

$$u_R^2 = v_x^2 + v_y^2 - 2r_{xy}v_x v_y \tag{6}$$

in which r_{xy} is the correlation coefficient between trip-level boardings and passenger-kilometers.

For the third term of Equation 5, the covariance is nonzero because both estimates depend on the ride check sample. Write

$$\text{Cov}(\bar{y}_{\text{I}}, R_{\text{I}}\bar{x}_{\text{II}}) = E\left[\bar{y}_{\text{I}}\left(\frac{\bar{y}_{\text{I}}}{\bar{x}_{\text{I}}}\right)\bar{x}_{\text{II}}\right] - E(\bar{y}_{\text{I}})\ E(R_{\text{I}}\bar{x}_{\text{II}})$$

$$= \bar{X}E\left(\frac{\bar{y}_{\text{I}}^2}{\bar{x}_{\text{I}}}\right) - \bar{Y}^2 \tag{7}$$

Conditioning on the Sample I mean boardings, write

$$E\left(\frac{\bar{y}_{\text{I}}^2}{\bar{x}_{\text{I}}} \mid \bar{x}_{\text{I}}\right) = \frac{1}{\bar{x}_{\text{I}}}\left\{\left[E(\bar{y}_{\text{I}} \mid \bar{x}_{\text{I}})\right]^2 + V(\bar{y}_{\text{I}} \mid \bar{x}_{\text{I}})\right\}$$

The second term inside braces is simply the variance of the ratio estimator, scaled by a constant. Assuming that the relative variance of the ratio is constant,

$$= \frac{1}{\bar{x}_{\text{I}}}\left[E(\bar{y}_{\text{I}} \mid \bar{x}_{\text{I}})\right]^2\left(1 + \frac{u_R^2}{n_{\text{I}}}\right) = R^2\bar{x}_{\text{I}}\left(1 + \frac{u_R^2}{n_{\text{I}}}\right)$$

Now dropping the condition by taking expectation over all possible Sample I boardings,

$$E\left(\frac{\bar{y}_{\text{I}}^2}{n_{\text{I}}}\right) = R^2\bar{X}\left(1 + \frac{u_R^2}{n_{\text{I}}}\right)$$

and so the covariance term (Equation 7) reduces to

$$\text{Cov}(\bar{y}_{\text{I}}, R_{\text{I}}\bar{x}_{\text{II}}) = \bar{Y}^2\frac{u_R^2}{n_{\text{I}}}$$

Combining all three components of Equation 5 and dividing by the squared mean of Y to get a relative variance (squared coefficient of variation),

$$v^2(\hat{y}_3) = (1 - w)^2\frac{v_y^2}{n_{\text{I}}} + w^2\left(\frac{v_x^2}{n_{\text{II}}} + \frac{u_R^2}{n_{\text{I}}}\right) + 2w(1 - w)\frac{u_R^2}{n_{\text{I}}} \tag{8}$$

Optimal Weights

As was pointed out earlier, the weight w is arbitrary in that the estimator is unbiased for any value of w between 0 and 1. In the interest of sampling efficiency, w can be chosen so that it minimizes the

variance of the estimator. Taking the derivative of Equation 8 with respect to w yields

$$w_{opt} = \frac{\frac{v_y^2}{n_I} - \frac{u_R^2}{n_I}}{\frac{v_y^2}{n_I} - \frac{u_R^2}{n_I} + \frac{v_x^2}{n_{II}}}$$

or, substituting with Equation 7,

$$w_{opt} = \frac{2r_{xy}v_y - v_x}{2r_{xy}v_y - v_x\left(1 - \frac{n_I}{n_{II}}\right)} \quad (9)$$

The optimum is rather flat, meaning that values of w near the optimal perform almost as well as the optimal, so that there is no practical penalty for choosing simple, rounded values for w. One intuitive way of selecting w is to examine an estimator based on expanding the ride check ratio by a combined estimator of the mean of x:

$$\hat{y}_4 = R_I\bar{x}_{I,II} = \left(\frac{\bar{y}_I}{\bar{x}_I}\right)\frac{\bar{x}_I n_I + \bar{x}_{II}n_{II}}{n_I + n_{II}}$$

This estimator is equivalent to that given in Equation 4, with weight

$$w = \frac{n_{II}}{n_I + n_{II}} \quad (10)$$

which can serve as a heuristic value for w. It may be noted that the heuristic weight equals the optimal weight when $v_x = r_{xy}v_y$, which is not far from true for typical transit systems.

Application to NFTA Bus and Light-Rail System

Mixed estimator sampling requirements were determined for both NFTA's bus and light-rail systems. The number of TPMS trips providing boardings counts was 182 for bus and 104 for light rail. To get an indication of the effectiveness of the mixed estimator, the question is how many ride checks would be necessary to achieve the NTD accuracy goal using a mixed estimator that takes advantage of the TPMS data versus the number of ride checks using simple expansion. The numerical analysis, including key statistical parameters, is given in Table 2. A summary of the results follows.

For the bus system, little is gained using the mixed estimator because the ride check sample requirement drops by only 6 percent (from 450 to 419 trips). The small gain is due to the rather weak correlation between boardings and passenger-kilometers (0.59), which makes extra boardings information of little value in estimating average passenger-kilometers per trip. The weak correlation between boardings and passenger-kilometers is primarily due to large differences between routes in average (passenger) trip length; that is, there are some routes (short routes) where the average trip length is small, and others (long, express routes) where average trip length is large.

On the other hand, using a mixed estimator yields a significant gain for the rail system, with the ride check sampling requirement falling by 38 percent (from 197 to 123). Unlike the bus system, the light-rail system consists of a single line, so there is no between-route variation in average trip length, only between-trip variation (which is typically considerably smaller), resulting in a strong correlation between boardings and passenger-kilometers (0.87).

TABLE 2 Analysis of Mixed Estimator Versus Simple Expansion for NFTA Bus and Light-Rail Systems

	Bus (many routes)	*Light Rail (one line)*
Statistical Parameters		
cv boardings	0.782	0.580
cv pass-km	1.082	0.715
correlation coefficient	0.589	0.874
unit cv of ratio	0.886	0.350
n_{II} (boardings counts)	182	104
Mixed Estimator Results		
optimal w	0.215	0.494
necessary n_I (ride checks)	**419**	**123**
resulting precision	9.98%	9.98%
Simple Expansion Results		
necessary n (ride checks)	**450**	**197**
resulting precision	9.99%	9.98%

On a practical note, sampling plans using the mixed estimator were developed for both NFTA's bus and light-rail system using greater sample sizes than those shown to provide a margin of safety with respect to the statistical parameters. The sampling plans also use cluster sampling to improve their cost-effectiveness. The improvements in efficiency from the mixed estimator, together with those from cluster sampling, enabled the development of a plan for obtaining both the TPMS sample and the necessary NTD sample using 23 percent fewer checker hours than would have been needed for a plan based on simple expansion and single-trip sampling.

Although it is dangerous to generalize from these two examples, it appears that the mixed estimator can be a valuable strategy in a setting that satisfies the following three conditions:

1. Total passenger boardings are not known but must be estimated through sampling;
2. A sample of boardings counts will be available or is useful for other purposes; and
3. The system has little variation in average passenger trip length because it consists either of a single route or of a group of routes that are similar in length and in express-local orientation.

SYMMETRY-BASED ESTIMATOR USING BOARDINGS DATA ONLY

Port Authority Transit in metropolitan Pittsburgh operates two heavily traveled light-rail lines, 42L and 42S, which share a common trunk extending 13 km south from the central business district (CBD). The total lengths of the two routes are 22 and 17 km, respectively. Conducting the TPMS passenger survey on these lines is particularly labor-intensive. In order to get good control totals (boardings) and give each boarding passenger a questionnaire, a surveyor is needed at each of three doors for approximately 48 h a year. Because stops on the light-rail lines are well known and spaced farther apart than stops on most bus lines, the surveyors can record boardings by stop. However, it is impractical to have them record alightings by stop.

Good estimates of boardings on these lines are available from other data collection efforts. Estimating passenger-kilometers therefore calls for determining average trip length, which normally requires data on both boardings and alightings. However, if passenger travel patterns over the day are symmetric, the boarding pattern in one direction should be the same as the alighting pattern in the opposite direction. A recent study done at Northeastern University by Navick and Furth (unpublished data) explores this assumption with the hope of estimating passenger-kilometers from enhanced farebox data. Full-day ride check data from five Los Angeles area bus routes were analyzed, with the result that differences in boarding and alighting patterns in opposite directions were not practically significant on most of the routes. If the symmetry assumption also holds on Pittsburgh's light-rail lines, it should be possible to estimate average trip length from boardings data only.

Derivation of Symmetry Estimator

A derivation of the symmetry estimator of average trip length, using boardings data only, follows. It takes a different approach than the Navick and Furth study, one that better lends itself to deriving an estimate from a sample. For convenience, let the downtown end of the line be specified as a reference point. Consider a single passenger, passenger *j*, traveling inbound on trip *i*. Let b_{ij} represent the location of his or her boarding stop, measured as the distance from the boarding stop to the reference point. Similarly, let a_{ij} represent the location of the alighting stop. The distance traveled by this passenger is therefore $(b_{ij} - a_{ij})$. Summing over all passengers in that direction yields total passenger-kilometers and dividing by the total number of passengers gives the average trip length (ATL):

$$\text{ATL} = \frac{1}{N}\sum_{i,j}\left(b_{ij} - a_{ij}\right) = \frac{1}{N}\sum_{i,j} b_{ij} - \frac{1}{N}\sum_{i,j} a_{ij} = \bar{b} - \bar{a} \tag{11}$$

where N is total number of passengers on that line in that direction. Equation 11 gives the interesting result that the average trip length is simply the difference between the mean boarding location, which may be called the boardings centroid, and the alightings centroid. With trip-level boardings data, which can be considered cluster sampling of passengers, the boardings centroid is estimated by

$$\bar{b} = \frac{\sum_{i}^{n} \bar{b}_i x_i}{\sum_{i}^{n} x_i} \tag{12}$$

where n is number of sampled trips, x_i is number of boardings on trip i, and the boardings centroid on trip i is

$$\bar{b}_i = \frac{1}{x_i}\sum_{j=1}^{x_i} b_{ij} \tag{13}$$

Following the symmetry assumption, the overall alightings centroid is equal to the boardings centroid in the opposite direction of travel (with locations measured from the same reference point), and average trip length is the same in both directions. Thus, a sample of trips in both directions with boardings recorded by stop provides a means of estimating average trip length, which can then be expanded by total route boardings to yield an estimate of total passenger-kilometers on the route. The procedure is as follows:

1. Define the location of each stop as its distance from a reference point (one end of the route). The same reference point (e.g., the downtown end of the route) must be used for both directions.
2. For each trip, calculate the trip level boardings centroid (Equation 13).
3. For each direction, aggregate over all trips to find the overall boardings centroid (Equation 12).
4. Average trip length is the absolute difference between the two.

The following section, in which a method for determining the variance of the symmetry estimate is derived, may be skipped without loss of continuity.

Variance of Symmetry Estimator

There are two potential sources of error in using the symmetry estimate: sampling error and modeling error, which occurs if the symmetry assumption is not true. Sampling error arises from the fact that the boardings centroids are estimated from a limited sample of trips. Assuming independence (a reasonable assumption, even if with round trip sampling, because inbound and outbound patterns at the same time of day are unrelated), the relative sampling variance of the centroid-based estimator is

$$v^2 = \frac{1}{ATL^2}\left(\frac{S_{b1}^2}{n_1} + \frac{S_{b2}^2}{n_2}\right) \tag{14}$$

where n_1 and n_2 are the number of trips sampled in Directions 1 and 2 and

S_{b1}^2, S_{b2}^2 = variance of the boardings centroid in direction 1, 2

The variance of the boardings centroids should be determined following formulas for cluster sampling with the ratio-to-size estimator (*9*). For Direction 1,

$$S_{b1}^2 = \frac{1}{n_1(\bar{x})^2}\sum_{i=1}^{n_1} x_i^2\left(\bar{b}_i - \bar{b}\right)^2 \tag{15}$$

with only trips in Direction 1 included in the sum. An analogous equation holds for Direction 2.

Normally the sample sizes in the two directions will be very nearly equal; if n is total number of sampled trips, n_1 and n_2 will each equal $n/2$. In that case, the square of the unit cv (found by removing n from Equation 14) becomes

$$cv^2 = \frac{2}{ATL^2}\left(S_{b1}^2 + S_{b2}^2\right) \tag{16}$$

Modeling Error

The precision of the estimate depends not only on sampling error but also on modeling error. Modeling error can be roughly estimated from a historic data set of ride checks. The ATL for this set of trips is known. Assuming an equal number of passengers in both

directions and that the reference end of the route is at the alighting end of Direction 1, and suppressing the notation for grand mean, ATL is

$$ATL_{\text{true}} = \frac{1}{2}[b_1 - a_1 - (b_2 - a_2)]$$

and the symmetry estimate is

$$ATL_{\text{sym}} = \frac{1}{2}[b_1 - \hat{a}_1 - (b_2 - \hat{a}_2)]$$

where the modifier ^ has the usual meaning "estimator of." The estimator of a_1 is b_2, which may be considered a sample estimate with variance S_{b2}^2/n_2, and analogously for the estimator of b_1. Therefore the standard error of ATL_{sym} is

$$S(ATL_{\text{sym}}) = \sqrt{\frac{1}{4}\left(\frac{S_{b1}^2}{n_1} + \frac{S_{b2}^2}{n_2}\right)} \tag{17}$$

and its mean-squared error is

$$MSE(ATL_{\text{sym}}) = E\left[(ATL_{\text{sym}} - ATL_{\text{true}})^2\right]$$
$$= S^2(ATL_{\text{sym}}) + e^2 ATL_{\text{true}}^2 \tag{18}$$

where e is the relative modeling error. An estimate of the relative squared modeling error can be found by replacing the $E[\ ]$ term of Equation 18 with the value obtained from the sample:

$$e^2 = \max\left\{\frac{1}{ATL_{\text{true}}^2}\left[(ATL_{\text{sym}} - ATL_{\text{true}})^2 - S^2(ATL_{\text{sym}})\right], 0\right\} \tag{19}$$

If the quantity in square brackets is negative, it means that the true and the estimated ATL differ by less than what would be expected because of sampling error, and therefore the modeling error may be safely neglected. If the quantity in square brackets is positive, the modeling error may still be zero, because the greater-than-expected deviation could just be a case of larger-than-expected sampling error. Under conventional hypothesis testing, in which the null hypothesis is that symmetry holds, the modeling error would be accepted as being zero at the 5 percent significance level as long as the squared difference between ATL measures was less than twice the variance of the estimate. Absent that prejudice toward the symmetry assumption, the best estimate of the squared modeling error is given by Equation 19. It should be emphasized, however, that this is a rough estimate based on a rather limited set of data.

Under the assumption that modeling error could lead as easily to overestimation as underestimation (an assumption that cannot be tested without a larger and more varied data set), the modeling error can be considered as an additional variance term that does not diminish with sample size. The precision of an estimate of passenger-kilometers based on the symmetry method can then be expressed in terms of the relative root mean squared error, given by

$$rmse = \sqrt{\frac{cv_{\text{sym}}^2}{n} + e^2}$$

and the precision of the estimate will simply be

precision = t (rmse)

where t, as usual, is the t-value associated with the specified confidence level. The number of degrees of freedom is hard to specify because of the inclusion of the modeling error term, but it is reasonable to use $n - 2$ as the number of degrees of freedom, since one degree of freedom is lost in estimating each of the S_{b2} terms.

Results for Symmetry Estimator

The symmetry estimator was analyzed for both lines 42L and 42S using data from July 1995 to June 1997. Some of the data had both ons and offs by stop; some had only ons by stop. The CBD (northern) end of the line is used as reference point. The results are shown in Table 3. Key aspects are the following.

TABLE 3 Analysis of Symmetry Estimator for Pittsburgh Light Rail

	Route 42L			Route 42S			42L, Excluding Downtown Trips	
a. Centroid	*mean (km)*	*std dev (km)*	*n*	*mean (km)*	*std dev (km)*	*n*	*mean (km)*	*std dev (km)*
Boardings In	12.7	4.8	29	9.8	4.9	33	14.4	2.7
Alightings Out	12.6	2.2	12	9.8	3.2	17	14.1	1.7
Boardings Out	2.2	1.9	29	2.2	1.0	33	2.3	2.2
Alightings In	2.1	1.4	12	1.1	0.5	17	2.4	2.1

b. Average Trip Length	*mean (km)*	*unit cv*	*std err (km)*	*mean (km)*	*unit cv*	*std err (km)*	*mean (km)*	*unit cv*
Inbound	10.6			8.8			12.0	
Outbound	10.4			7.6			11.7	
Combined	10.5	0.30		8.2	0.34		11.9	0.24
Symmetry estimate	10.5	0.70	0.48	7.7	0.91	0.43	12.1	0.40
$n_{\text{sym}} / n_{\text{ratio}}$		5.4			7.2			2.7
Relative modeling error		0.0%			3.2%			0.0%

First, regarding whether symmetry holds or not, the results are mixed. The two routes afford four centroid comparisons, one for each route and direction. Three of the four are very close. The fourth, with a discrepancy of 1.1 km, occurs on Route 42S, where inbound alightings are more heavily concentrated toward the CBD end of the line (centroid = 1.1 km) than are outbound boardings (centroid = 2.2 km). Averaging over both directions on a route, the symmetry estimate for ATL on Route 42L matches the true value (for the set of sampled trips) almost exactly (10.5 km). However, on Route 42S, the discrepancy between the centroids at the CBD end leads to a difference between the two ATL measures of 0.5 km, more than can be explained by sampling error alone. The relative modeling error for Route 42S is estimated to be 3.2 percent. However, it should be noted that for both routes the alightings centroids are estimated from small samples (12 and 17 trips, respectively), and therefore the results must be regarded as somewhat tentative.

Second, regarding between-trip variability of the boardings centroids, the results are again mixed. In the outbound direction, the standard deviation is 1.9 km and 1.0 km on Routes 42L and 42S, respectively. However, in the inbound direction for both lines, the standard deviation of the boardings centroid is about 4.8 km, which is about half the average trip length. This very high level of variability can be explained in part by two phenomena. One is that at a pair of transfer points only about 3 km from the CBD end of the route (and thus outside the primary boarding area), the number of passengers boarding on a.m. peak trips fluctuates wildly, for example, from 0 on one trip to 50 on the next, because of connections with other transit routes. In the p.m. peak the number of passengers alighting at these stations is far less variable and smaller as well. A second reason is that in the a.m. peak some trains serve the trunk only. Naturally, those trips have a boardings centroid that is closer to the CBD; at the same time, they distort the boardings pattern for the following trips, whose boardings centroid moves farther from the CBD.

The large variation in the inbound boardings centroid leads to a high unit (i.e., per trip) *cv* for the symmetry estimate of average trip length. On Route 42L, the unit *cv* is 0.70, which compares unfavorably with the ratio estimate's unit *cv* of 0.30. (The ratio estimator is the method used with ride check data when annual boardings are known.) As necessary sample size is inversely proportional to the square of the unit *cv*, this means that the symmetry estimator would require sampling 5.4 times as many trips as would the ratio estimator. Although the symmetry samples require counting only boardings by stop, whereas the ratio samples require counting both boardings and alightings by stop, this difference is not enough to compensate for a sample five times as large. The effect is even stronger on Route 42S.

To get an idea of the applicability of the symmetry estimate to other transit lines, the analysis for Route 42L was repeated with passengers whose entire trip lies within 4 km of the CBD end of the line excluded. The result, also given in Table 3, is a large decrease in the standard deviation of the inbound boardings centroid and a decrease in the unit *cv* of average trip length to 0.40. Comparing with the unit *cv* of the ratio estimate (which drops to 0.24 with downtown passengers excluded), the symmetry estimator still needs 2.7 times as many trips sampled, with boardings recorded only, as the ratio estimator needs with both boardings and alightings recorded by stop. If other causes of sharp variability (e.g., short-turning trips) were also absent, that factor would likely be smaller still. Depending on the number of boardings-only counts needed for other data collection efforts and the relative cost of a ride check versus a boardings count, the symmetry estimator may be useful in some contexts.

CONCLUSIONS

When additional data, be they from electronic fareboxes or boardings counts done for another survey, are available, sampling plans to estimate passenger-kilometers can often be developed that are more efficient than simply expanding a sample mean or following the default NTD sampling plan. Three sampling plans are described that were developed for different contexts. Variance formulas are given so that others may use them, and numerical results from U.S. transit systems are presented to give an idea of the value of the three methods.

In the first context, a small bus system with complete boardings data, a stratified sampling method called combined ratio estimation was applied and shown to be very efficient. This method is applicable in transit systems in which total boardings are known and is likely to be effective for systems with up to 50 routes.

The second context is a transit system without information on total boardings but with a sample of boardings counts available. An estimator that mixes ride check data and boardings counts was found to lead to sizable reductions in the number of ride checks needed for the rail system, which consists of a single line, but not for the bus system, which has a large number of routes with widely varying route lengths.

The third context is a light-rail line with known total boardings and available counts of boardings by stop but not alightings by stop. Data show mixed support for an assumption of symmetry in boarding and alighting patterns in opposite directions of travel. Moreover, the between-trip variance in boardings centroid on the routes studied was too great to make this method of estimation effective compared with other methods. Nevertheless, it may have promise in other contexts where boarding patterns do not fluctuate as much between trips.

ACKNOWLEDGMENTS

This research was supported in part by the Federal Transit Administration and the American Public Transit Association. The efforts of the staff of the transit agencies involved in providing data are gratefully acknowledged.

REFERENCES

1. *Sampling Procedures for Obtaining Fixed Route Bus Operation Data Required Under the Section 15 Reporting System.* Circular 2710.1A. FTA, U.S. Department of Transportation, 1990.
2. *Revenue-Based Sampling Procedures for Obtaining Fixed Route Bus Operation Data Required Under the Section 15 Reporting System.* Circular 2710.4. FTA, U.S. Department of Transportation, 1985.
3. Phifer, S. P. Use of Sampling in Bus-Line Data Collection. In *Transportation Research Record 877,* TRB, National Research Council, Washington, D.C., 1982, pp. 41–44.
4. Furth, P. G., J. P. Attanucci, I. Burns, and N. H. Wilson. *Transit Data Collection Design Manual.* Report DOT-I-85-38. U.S. Department of Transportation, 1985.

5. Furth, P. G., and B. McCollom. Using Conversion Factors to Lower Transit Data Collection Costs. In *Transportation Research Record 1144,* TRB, National Research Council, Washington, D.C., 1987, pp. 1–6.
6. Furth, P. G., K. L. Killough, and G. F. Ruprecht. Cluster Sampling Techniques for Estimating Transit System Patronage. In *Transportation Research Record 1165,* TRB, National Research Council, Washington, D.C., 1988, pp. 105–114.
7. Huang, W. J., and R. L. Smith. Development of Cost-Effective Sampling Plans for Section 15 and Operational Planning Ride Checks: Case Study for Madison, Wisconsin. In *Transportation Research Record 1402,* TRB, National Research Council, Washington, D.C., 1993, pp. 82–89.
8. Furth, P. G., and A. Kumar. Ridership Sampling for Barrier-Free Light Rail. In *Transportation Research Record 1402,* TRB, National Research Council, Washington, D.C., 1993, pp. 90–97.
9. Cochran, W. G. *Sampling Techniques,* 3rd ed. John Wiley and Sons, Inc., New York, 1977.

Publication of this paper sponsored by Committee on Transit Management and Performance.