U.S. Department of Transportation
Urban Mass Transportation Administration
and Transportation Systems Center

Statistical Issues Affecting Transit Data Collection: Analysis of Pittsburgh Ride Check Data

Final Report

July 1986

Multisystems, Inc.

1050 Massachusetts Avenue
Cambridge, Massachusetts 02138

Service and Methods Demonstration Program

U.S. Department of Transportation
Urban Mass Transportation Administration
and Transportation Systems Center

Statistical Issues Affecting Transit Data Collection: Analysis of Pittsburgh Ride Check Data

Final Report

July 1986

Multisystems, Inc.

1050 Massachusetts Avenue
Cambridge, Massachusetts 02138

Service and Methods Demonstration Program

Introduction[1]

In fall 1983, spring 1984, and fall 1984, PAT (the Pittsburgh regional transit authority) conducted intensive ride checks on selected routes. Data from these ride checks were used to explore issues that remained unresolved by the Bus Transit Monitoring Study, whose final product was the Transit Data Collection Design Manual (TDCDM)[2]. This report is divided into sections, each covering one of the following areas:

1. Multi-Year Use of Conversion Factors

2. Optimal Sizing of Baseline Sample for Conversion Factors: More Flexibility than TDCDM

3. Effect of Cluster Sampling

4. Default Formulas for Estimating Correlation

5. Evaluation of PAT's Proposed Approach toward Meeting Section 15 Requirements

---

1. The author is Peter G. Furth. Valuable technical assistance from Jim Wensley and the cooperation of Rich Feder of PAT are gratefully acknowledged.

2. P. G. Furth et. al., Transit Data Collection Design Manual, Final Report DOT-I-85-38, UMTA, June, 1985.

1. MULTI-YEAR USE OF CONVERSION FACTORS

Problem Description

Monitoring of transit data can be made much less expensive through the use of conversion factors. A conversion factor is a ratio between the means of two data items, an "inferred item" (in the numerator) and an "auxiliary item" (denominator). By measuring the mean of the auxiliary item and multiplying it by the conversion factor, the mean of the inferred item can be estimated. Usually the auxiliary item is chosen to be an item that is easily collected and that bears a stable statistical relationship to other desired items that are more difficult or expensive to measure. In this way, one need only measure one data item to get estimates of several.

In the classical application of conversion factors (better known as "ratio estimates"), the conversion factor and the mean of the auxiliary item are measured in the same time frame. A sample of size n_2 of the auxiliary item is measured, and for a randomly selected subset of n_1 observations, the auxiliary item is measured simultaneously. From the paired sample of size n_1, the conversion factor is estimated, and from the sample of size n_2 the mean of the auxiliary item is measured. Their product is the estimated mean of the inferred item. The accuracy of this product depends on both n_1 (which determines the accuracy of the conversion factor) and on n_2 (which determines the accuracy of the auxiliary item's mean). Minimizing data collection costs to achieve a desired accuracy for the inferred item (or for both the auxiliary and inferred item) yields a formula that balances the cost and resulting accuracy of the paired sample with that of the auxiliary item sample. This minimum cost can then be compared with the cost of directly estimating the inferred item (i.e. without a conversion factor) to determine whether the conversion strategy should be used.

The approach taken in the Transit Data Collection Design Manual (TDCDM) differs from this classical approach in that the conversion factor and the mean of the auxiliary item are estimated in different time frames (and therefore with non-overlapping samples). As suggested by TDCDM, a paired sample of size n_1 is taken in a "baseline year" (year 0), yielding a conversion factor, and then the auxiliary item is measured with a sample of size n_2 in each of the "monitoring years" 1, 2, ..., yielding a current estimate of the mean of the auxiliary item which, when multiplied by the conversion factor, yields a current estimate of the mean of the inferred item. Given the number of years one plans to use the conversion factor (before reestimating it with a paired sample) and other pertinent parameters, the total cost of data collection, encompassing the baseline year and the monitoring years, can be minimized, yielding the formula given in TDCDM.

However, the ratio of the inferred to auxiliary item (e.g. of boardings to load) is not guaranteed to be unchanging over the years. Therefore, as the number of years since the baseline year increases, the reliability of the conversion factor estimated in that baseline years becomes more doubtful. In recognition of this phenomenon, TDCDM makes this recommendation: that conversion factors be reestimated whenever there is a significant change in the route, the level of service offered, or in the demand. Demand will usually be monitored regularly by some passenger-use related item, and TDCDM recommends reestimating conversion factors when this item changes by 25% from its baseline level. No explicit maximum life of a conversion factor is spelled out, although we feel that in no case should a conversion factor be used for more that 5 years without being reestimated.

The meager recognition that TDCDM gives to the problem of the conversion factor that loses its precision over time prompted this study. Two methods of dealing with this loss of precision are studied. The first is a test of the 25% change threshhold for the monitored passenger use item. The second is a model to explicitly account for this growing uncertainty.

The data used in this study is a set of ride checks performed on selected routes in Pittsburgh. The data were supplied by the local transit agency, PAT. Checks were made in Fall '83, Spring '84, and in Fall '84, with roughly the same routes checked each time. Sample sizes vary from route to route in each dataset. Our study uses load at the peak load point as the auxiliary item, and boardings as the inferred item. The data permit analysis of a 1-year change at most, so that some questions cannot be adequately answered. The data also permit some testing of seasonal change of conversion factors.

Relationship of Change in Conversion Factor to Change in Passenger Demand

Our first analysis was to test the hypothesis that R (the conversion factor) changed between Fall '83 and Fall '84 on the R/D/TP's (route/direction/time period combinations) that are in both the Fall datasets. R/D/TP's were also classified according to how much peak load had changed in that year. Our goal was to see whether R changed more frequently on R/D/TP's that experienced greater changes in peak load.

The test used was the standard difference of means test. The conditions for the test are not met exactly, as R (the measured estimate of R) is not normally distributed. However, its distribution is close to normal, and by virtue of the Central Limit Theorem the difference in R's between the two years is still closer to normal. The test variable is

$$D = R_1 - R_2$$

where R_1 and R_2 are the estimated convertion factors in Fall '83 and Fall '84 respectively. Under the null hypothesis H_o, the mean of D is 0 and its estimated variance is $S_d^2 = \text{Var}(R_1) + \text{Var}(R_2)$. The estimated variances of the conversion factor estimates were calculated using the formula given in TDCDM. For a given confidence level c, H_o is rejected if D is too far from 0, that is,

$$\text{Reject } H_o \text{ if } \quad D > s_d t_c$$

where t_c it the critical t-value for the confidence level c. The confidence level was set at 90%, and of the 61 R/D/TP's common to the Fall datasets, H_o was rejected for 12, or 20% of them.

Table 1 shows the incidence of acceptance/rejection of H_o for R/D/TP's classified by percentage change in peak load. Of the R/D/TP's experiencing a change in peak load of $\pm$25% or more, H_o is rejected for 17%, while H_o was rejected for 20% of the R/D/TP's with a smaller relative change in peak load. In this case, then, large changes in peak load do not indicate a greater incidence of change in R. However, if the threshhold for change in peak load is altered to be $\pm$20%, the opposite result occurs: H_o is rejected for 33% of the R/D/TP's with a change of $\pm$20% or more in peak load, compared to 15% for the R/D/TP's with less change. Taken together, these results offer weak support, if any, of the idea that big changes in monitored demand indicate significant changes in R.

We also explored the possibility that changes in R are more related to absolute (rather than percentage) changes in peak load. The results are shown in Table 2. They show at best a very weak relationship between absolute change in demand and change in R. The threshhold yielding the sharpest difference is $\pm$4 passengers per trip; below that threshhold, H_o is rejected on 12% of the routes, while the rejection rate is 37% above it. But threshholds are rather arbitrary, and not easily transferable to other systems, and with a different threshhold, the difference in rejection rates is small. So while rejection rates seem better related to absolute change in demand than to percentage change, the relationship still seems to weak to justify differing treatment of routes with little change vs. greater change.

The same analysis was done to test for changes between Fall '83 and Spring '84. Of the 101 R/D/TP's found in both datasets, 22, or about 22%, showed significant change in R. When broken down according to percentage and absolute change in peak load, the relationship is again very weak. If the threshholds that yielded the sharpest division ($\pm$20% and $\pm$4) are applied, virtually no difference in rejection rates is found between the sets above and below the threshhold. (The threshholds yielding the sharpest distinction for this case are $\pm$25% and $\pm$6.)

The results of this analysis indicate that change in monitored demand level is at best a very rough and inefficient indicator of change in R, both for 1-year and the seasonal changes. While there were no data to test the usefulness of this indicator for multi-year changes, there is no reason to expect the results to be any different.

Relationship Between R and Change in R

Another potential indicator of a changing conversion factor that was considered was the value of the conversion factor itself. This can perhaps best be explained by recognizing that the ratio (boardings)/(peak load) cannot be less that 1, and that therefore there is little room for change when R is near 1. The 61 R/D/TP's appearing in both Fall datasets were ordered by increasing value of R (as measured in Fall '83), and the rejection rates of H_o (the hypothesis that R changed between the two years, as described above) was noted for low, medium, and high values of R. The data, summarized in Table 3, show no correlation between R and the rejection rate.

Modeling the Increase in Uncertainty

This approach recognizes that the true conversion factor can change from year to year. If the true conversion factor is

μ_0 in the baseline year and is μ_t t years after the baseline year, the change in the true mean from year t-1 to year t can be called δ_t, as shown below:

$$\mu_1 = \mu_0 + \delta_1$$

$$\mu_2 = \mu_1 + \delta_2 = \mu_0 + \delta_1 + \delta_2$$

$$\mu_t = \mu_0 + \sum_{i=1}^{t} \delta_i = \mu_0 + D_t \qquad (1)$$

where
$$D_t = \sum_{i=1}^{t} \delta_i \qquad (2)$$

is the cumulative change over t years.

Because $E[R_0] = \mu_0 \neq \mu_t$ (for $t \neq 0$), R_0 is a "biased" estimator of μ_t. The bias is $-D_t$, since $E[R_0] = \mu_t - D_t$. In making predictions about a route in the absence of special information on how its conversion factor changes, this bias is unknown. There is no reason to believe that it is systematically positive or negative; in this sense, the baseline conversion factor is "neutral biased". Treating δ_i for a particular R/D/TP as a random variable, the following assumptions are made:

1) δ_i is independent of δ_j for $i \neq j$; i.e., the change in the true conversion factor is independent from year to year.

2) $E[\delta_i] = 0$, otherwise there would be an upward or downward trend in R over time.

3) $Var[\delta_i] = \sigma_\delta^2$ for all i (i.e. the variance is the same in every year).

From these assumptions it follows that

$$E[\delta_i^2] = Var[\delta_i] = \sigma_\delta^2$$

$$E[D_t] = 0; \quad Var[D_t] = E[D_t^2] = t\sigma_\delta^2$$

When using a biased estimator, instead of its variance the proper measure of its accuracy is its mean squared error (MSE), given by

$$\begin{aligned} MSE_t(R_0) &= E[(R_0 - \mu_t)^2] \\ &= E[(R_0 - \mu_0 + D_t)^2] \\ &= Var[R_0] + D_t^2 \end{aligned}$$

which is the well-known result that the MSE of an estimator is the variance of the estimator (about its own mean) plus the square of the bias. But since we are treating the bias itself as a random variable, we shall use the expected mean squared error, given by

$$\widehat{MSE}_t = E[MSE_t] = Var[R_0] + t\sigma_\delta^2 \qquad (3)$$

This expected MSE_t should now be used instead of the variance of R, and instead of v_R^2 (the squared coefficient of variation, or C.O.V., of R), the quantity that should be used is v_t^2, given by

$$v_t^2 = \widehat{MSE}_t \,/\, R_0^2 = v_R^2 + t\beta^2 \qquad (4)$$

where $\beta = \sigma_\delta / R_0$ is the relative standard deviation of the annual shift in the conversion factor.

The confidence interval for R is determined by the C.O.V. of the estimate of R. When using a biased estimator, it depends instead on v_t, which rises with t, reflecting the increasing uncertainty in where the current true mean is when the only measured statistic is the baseline year mean. Figure 1 illustrates how the confidence interval grows wider over time.

The sample size and accuracy formulas that depend on a baseline conversion factor should also use v_t^2 instead of v_R^2 (as TDCDM has). The effect of this change is to increase the monitoring phase sample size year by year to achieve a desired accuracy (because the conversion factor becomes less accurate, and so the monitoring sample must be more accurate to compensate). This formula becomes

$$n_2 = \frac{v_X^2 (1 + v_t^2)}{0.31\, d_m^2 - v_t^2} = \frac{v_X (1 + v_R^2 + t\beta^2)}{0.31\, d_m^2 - v_R^2 - t\beta^2} \qquad (5)$$

Equation (5) applies only if the denominator is greater than zero. A zero or negative denominator indicates that the desired accuracy cannot be achieved.

Equation (5) replaces TDCDM equation (6.15) (under this approach) for determining the monitoring phase sample size. Two remaining and related issues are determining how many years to use the conversion factor (before reestimating it) and the size n_1 of the baseline sample for estimating the conversion factor.

The mathematics involved in answering these questions is rather complex, which serves as a drawback to the approach. The answers cannot be expressed with a formula, and instead they were tabulated. For a given set of parameters, different values of n_1 are tried, beginning with the minimum allowed sample size of 10. For subsequent monitoring years 1,2,3,4, and 5, the necessary monitoring sample size n_2 is calculated from equation (5). Then the average yearly cost of data collection was calculated assuming the conversion was used 1,2,3,4, or 5 years. The combination of the value of n_1 and the number of years that yields the smallest average yearly cost is then the preferred stategy. To cover the expected range of parameter values, 40 such tables were generated, using two different values of β^2. These tables are shown in Appendix B. They show that as the desired tolerance in the monitoring phase becomes narrower and as the inherent variability in the data items increases, necessary sample size increases and the life of the conversion factor decreases.

These tables indicate that for the lower value of β^2, for tolerances down to $\pm$20%, the recommended life of the conversion factor is the full five years, and the baseline sample size is most often the same as the TDCDM baseline sample size (i.e. without adjustment for increasing uncertainty overtime). Likewise, with the larger value of β^2, little difference is found down to a tolerance of $\pm30\%$.

The use of these tables is illustrated with the same example found in Section 6.4.2 of TDCDM. The following parameters are given:

$v_X = 0.410$ = C.O.V. of peak load

$v_Y = 0.369$ = C.O.V. of boardings

$r_{XY} = 0.95$ = correlation coefficient

$d_m = 0.3$ = desired tolerance for boardings

service frequency = 5 trips per hour (12 min headway)

cycle time = 84 min

c = 0.28 = ratio of monitoring sample unit cost to baseline sample unit cost

$$C_{XY} = 0.017 = v_X^2 + v_Y^2 - 2r_{XY}v_Xv_Y$$
$$= .41^2 + .369^2 - 2(.95)(.41)(.369)$$

$$cv_X^2 = 0.048 = 0.28(.41)^2$$

The per unit monitoring sample cost c is explained as follows. In the monitoring phase, point checks will be used, requiring one checker-hour to observe 5 trips, or 0.2 checker-hrs per trip. (If the same checker observed both directions of travel, then only 0.1 checker-hrs per trip would be needed.) In the baseline phase, a checker is needed for each trip, and since each trip (one-way) takes 42 min, the cost is 42/60 = 0.7 checker-hrs per trip. Therefore the ratio c is 0.2/0.7 = 0.28.

The table chosen within Appendix B is Table (b), determined by the desired tolerance (0.3) and C_{XY} (near 0.015). An assumption about the value of β^2 must be made. For β^2 = 0.0024, enter the table with cv_X^2 (near 0.060), and find that the minimum cost stategy is to have a baseline sample of 10 trips and for the resulting conversion factor to have a life of 5 years. Assuming β^2 = .00064 yields the same result. Because the smallest permissible baseline sample size is 10 and the maximum recommended life of a conversion factor is 5 years, modeling the increasing uncertainty over time did not, in this case, affect the baseline sampling strategy. (It may still be that n_2 rises somewhat over the 5 year life of the conversion factor, in accordance with eq. (5).)

Estimating σ_δ^2

An estimator of σ_δ^2 is

$$s_\delta^2 = \frac{1}{r} \sum_{i=1}^{r} \left[(R_{1i} - R_{0i})^2 - s_{R1i}^2 - s_{R0i}^2 \right]$$

where r = number of R/D/TP combinations

R_{1i}, R_{0i} = estimated conversion factor for R/D/TP i in fall '84, fall '83

s_{R1i}^2, s_{R0i}^2 = sample variance of the estimated conversion factor for R/D/TP i in fall '84, fall '83

The proof is given in Appendix A. Basically, s_δ^2 compares the squared change in the estimated conversion factor with the change expected due to sampling error. If the true conversion factor underwent no annual change, then $E[s_\delta^2] = 0$.

On the other hand, if some (or all) of the conversion factors changed so that if the average squared change in the true conversion factor is $\sigma_\delta^2 = \frac{1}{r}\sum_{i=1}^{r}\delta_i^2$, then $E[s_\delta^2] = \sigma_\delta^2$. Then given this estimator of σ_δ^2 , an estimator of $\rho^2 = \sigma_\delta^2 / R^2$ is $h^2 = s_\delta^2 / \bar{R}^2$ where $\bar{R}$ is the average conversion factor for the sample.

We calculated s_δ^2 based on the 61 R/D/TP's with at least 10 paired observations in each of the Fall datasets. The results were s_δ^2 = .00106. The average R for the sample was $\bar{R}$ = 1.31, yielding the estimate b^2 = .00062.

We also classified the R/D/TP's according to the magnitude of R (from Fall '83). Because $b^2 = s_\delta^2 / \bar{R}^2$, a greater $\bar{R}$ should imply a smaller b^2 if s_δ^2 is the same for high and low R. Our results, shown in Table 3, indicate that s_δ^2 is <u>not</u> independent of R, but rises with R. In the group with low R ($R \leq 1.15$), s_δ^2 was slightly negative, indicating no change in estimated conversion factors for this group that could not be more than explained by sampling error along. For the middle group ($1.15 < R \leq 1.35$), s_δ^2 = .00064, and for the high group ($R > 1.35$), s_δ^2 = .0024. The increase in s_δ^2 as R increases more than compensates for an increase in $\bar{R}$, so that b^2 also increases with R. Our estimates are b^2 = 0 for the low category (since a negative value would be meaningless), b^2 = .00039 for the middle category, and b^2 = .0011 for the high category.

<u>Overall Recommendations</u>

Ideally, if cost did not matter, conversion factors would be estimated in every year. If this cannot be done, and the conversion factor estimated in a previous year is used, the method described above provides a way for compensating for the increased uncertainty. However, the drawbacks of this approach should be recognized. First, it will yield the desired accuracy only in an "average" sense. That is, while the annual change is more on some routes than on others, this approach models an equal change for all (within a category). Therefore, some routes will have greater accuracy than needed; other routes will have less accuracy. Second, there is very little basis for selecting a value of b^2. The Pittsburgh data cover one year's change only, apply only to conversion from load to boardings, and are not necessarily transferable to other systems. Finally, the calculations become more complex.

Our general recommendations on the multi-year use of conversion factors are as follow:

1.) New conversion factors should be estimated when a route undergoes a significant change. However, a change in measured demand alone should not force the reestimation of a conversion factor.

2.) The maximum life of a conversion factor should be 5 years after the baseline year.

3.) There is no need to apply the increasing uncertainty model when: (1) the desired tolerance is $\pm$30%, or (2) the desired tolerance is $\pm$20% and the conversion factor is in the range [0.75 to 1.35]. For other cases, ignoring the effects of increasing uncertainty will lead to systematically worse-than-desired accuracy.

4.) Because no good indicator of a change in conversion factor was found, regular estimation (i.e. annual or biennial) of conversion factors is recommended when the budget permits, until it can be determined that a conversion factor is stable. This approach could be applied to conversions that fail the criteria of recommendation (3) above as an alternative to applying the increasing uncertainty model.

The reader is reminded, however, that the above recommendations are tentative because the data examined covered only one city, one year's span, and one conversion (load to boardings).

Table 1

Change in R vs. Percentage Change in Average Peak Load

Percent change in average peak load

-25 -20 -15 -10 -5 0 5 10 15 20 25

													Total
Total#	2	1	5	4	10	13	6	3	3	2	2	10	61
# Changed	0	1	1	1	1	1	2	0	1	0	2	2	12

Table 2

Change in R vs. Numerical Change in Average Peak Load

Numerical change in average peak load

-10 -8 -6 -4 -2 0 2 4 6 8 10

													Total
Total#	1	3	0	4	9	18	8	7	4	4	1	2	61
# Changed	0	2	0	1	0	2	2	1	2	2	0	0	12

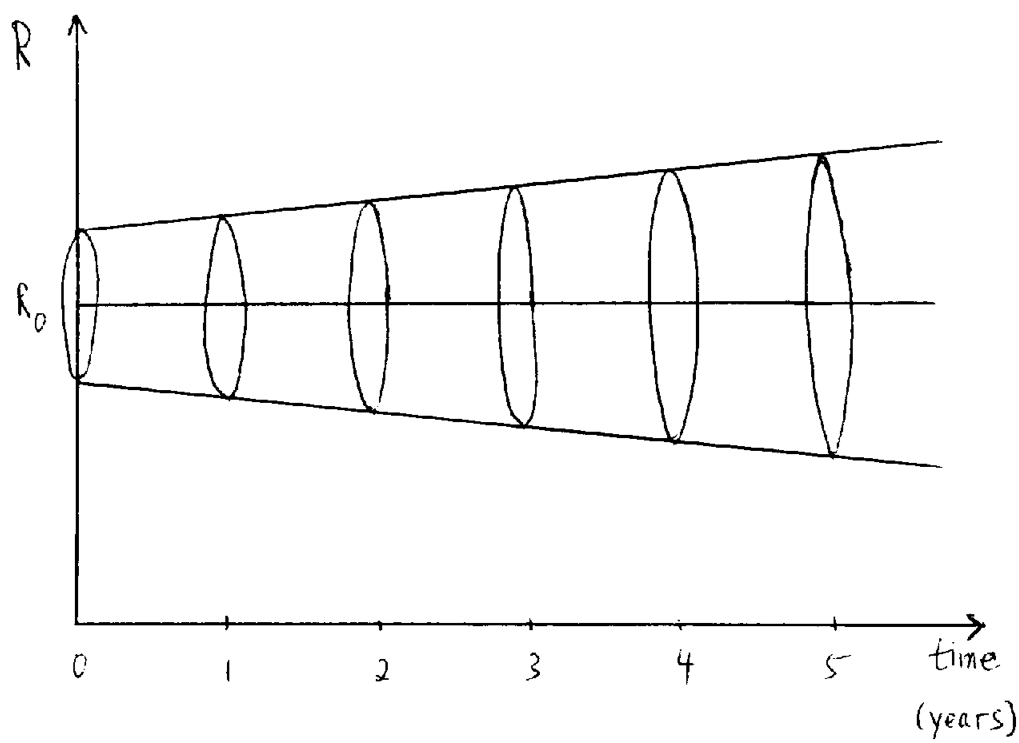
Table 3

Change in R vs. R

Range of R	1.00 - 1.15	1.16 - 1.35	1.36+	Total
Total #	18	20	22	61
# changed	4	3	5	12
% changed	22%	15%	23%	20%

Figure 1

Confidence Interval Growing With Time



2. EFFECT OF CLUSTER SAMPLING

Sample size formulas, both in the Transit Data Collection Design Manual (TDCDM) and elsehewere, typically assume simple random sampling. But because of the nature of transit service, it is far more convenient and inexpensive to sample in natural groups (or "clusters") than to sample the same number of unrelated, randomly selected units. Examples of natural clusters are:

<u>for ride checks</u>: cluster of trips = all the trips belonging to a run on a given day

<u>for point checks</u>: cluster of trips = all the trips passing the checkpoint in a given time period on a given day

<u>for passenger surveys:</u> cluster of individuals = all the individuals on a given vehicle or at a given transfer point at a given time on a given day

"Cluster sampling" is defined to mean random selection of clusters, and then sampling every element in the selected clusters. Thus, for ride checks, runs would be selected at random from a list of runs, and then every trip belonging to the run would be measured in the sample.

One way of comparing simple random sampling with cluster sampling is the quantity known as Kish's <u>deff</u> (for <u>design effect</u>), given by

$$\underline{\text{deff}} = \frac{\text{sample size under cluster sampling}}{\text{sample size under simple random sampling}}$$

Note that the numerator should be number of elements sampled (rather than the number of clusters sampled).

If <u>deff</u> > 1, cluster sampling will require that more elements be sampled than will random sampling. Nevertheless, the cost advantages of cluster sampling can make it the preferred alternative even if deff is significantly above 1.

Formulas for <u>deff</u> will be given for three types of estimates: means, ratios (i.e. conversion factors), and proportions. The following notation will be used, following Cochran (1977), whose text is the basis of the theory presented here:

N = number of clusters

M_i = size of cluster i

$\bar{M}$ = average cluster size

v_M^2 = (C.O.V.)^2 of cluster size = $\dfrac{\sum_{i=1}^{M}(M_i-\bar{M})^2}{(N-1)\ \bar{M}^2}$

$M_o = N\bar{M}$ = total number of elements in sample

Y = name of variable being measured

y_{ij} = value for the j'th observation in cluster i

$\bar{y}_i$ = mean per element in cluster i = $\frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij}$

$\bar{\bar{Y}}$ = overall mean per element = $\frac{1}{M_o}\sum\sum y_{ij}$

S^2 = overall variance of Y = $\left[\sum\sum(y_{ij}-\bar{\bar{Y}})^2\right]/(M_o-1)$

(CPS) = cross product sum = $2\sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{k=j+1}^{M_i}(y_{ij}-\bar{\bar{Y}})(y_{ik}-\bar{\bar{Y}})$

(NCPS) = normalized cross product sum = $\dfrac{(CPS)}{(N\bar{M}-1)S^2}$

ρ = intracluster correlation

$$= \frac{E[(y_{ij}-\bar{\bar{Y}})(y_{ik}-\bar{\bar{Y}})]}{S^2} \quad \text{for } j \neq k$$

$$= \frac{(NCPS)}{\bar{M}(1+v_M^2)-1}$$

The cross product sum (CPS) can be calculated another way which may be simpler in some situations:

$$(CPS) = \sum_{i=1}^{N}(\bar{y}_i-\bar{\bar{Y}})^2 M_i(M_i-1) \; - \; \sum_{i=1}^{N} M_i\, s_{wi}^2$$

where S_{wi}^2 is the variance of Y within cluster i, i.e.

$$s_{wi}^2 = \frac{1}{M_i}\sum_{j=1}^{M_i}(y_{ij}-\bar{y}_i)^2$$

(Notice that CPS entails the "between cluster" and "within cluster" elements of the first BTMS manual.)

Cluster Sampling for Means

If the item being estimated is $\bar{\bar{Y}}$, the mean per element (or the grand total, $M_0 \bar{\bar{Y}}$), then

$$\underline{\text{deff}} = 1 + (\text{NCPS}) \tag{1a}$$

$$= 1 + \rho \left[\bar{M}(1 + v_M^2) - 1 \right] \tag{1b}$$

Note that deff can be larger or smaller than 1.0, depending on the sign of the intracluster correlation ρ. If clusters tend either to consist primarily of large values of Y or to consist primarily of small values of Y, ρ will be positive, while if clusters tend to have an even mix of high and low Y ρ will be negative. In the former case ($\rho > 0$), cluster sampling demands that more elements be sampled, while in the latter case ($\rho < 0$), cluster sampling calls for fewer elements. Factors that tend to raise ρ in different situations are: greater ridership on some days than others, one route or run having generally higher ridership than another, and homogenaity of the population aboard a vehicle during an on-board survey. Factors that tend to lower ρ are: low load trips following high load trips due to "bunching", runs including trips that span (within a time period) times of greater and lesser utilization, and long radial trips whose population is a cross-section of society at large. Overall, we expect the positive-contributing factors to dominate in most applications, leading to larger samples than those demanded by simple random sampling.

Cluster Sampling for Category Proportions

Here the variable of interest is

P = proportion of the population in a given category

Define

P_i = proportion of the cluster i population in the given category

Then
$$\rho = \frac{\Sigma \rho_i M_i}{N \bar{M}}$$

Then
$$\underline{\text{deff}} = \frac{\Sigma (\rho_i - \rho)^2 M_i^2}{N \bar{M} \rho (1-\rho)} \qquad (2)$$

Equation (2) is equivalent to equations (1a) and (1b) if Y is defined to be 1 if the individual belongs to the given category and zero otherwise. Then ρ will be positive, implying <u>deff</u> $>$ 1, if individuals in the given category tend to be clustered together while individual outside the category are also clustered (but in different clusters). Conversely, ρ will be negative and <u>deff</u> $<$ 1 if clusters tend to be representative of the population distribution.

<u>Cluster Sampling for Ratio Estimates</u>

Here the variable of interest is the ratio

$$R = \bar{\bar{Y}}/\bar{\bar{X}}$$

where $\bar{\bar{Y}}$ and $\bar{\bar{X}}$ are the overall mean per element of the inferred and auxiliary variables, respectively. (The sample is a paired sample, with a value of both X and Y being recorded at each observation). Define the variable

$$d_{ij} = y_{ij} - Rx_{ij}$$

Because $R = \bar{\bar{Y}}/\bar{\bar{X}}$, the overall mean of d_{ij} is 0. Then if d_{ij} is substituted for y_{ij} as the variable of interest, equation (1a) yields the design effect <u>deff</u> for the ratio estimate

$$\underline{\text{deff}} = 1 + (\text{NCPS})_d \qquad (3)$$

where $(\text{NCPS})_d$ is the normalized cross product sum for the variable d. Applying the definitions given earlier and the relationship $\bar{\mathfrak{d}} = 0$, we get

(4)

$$\underline{deff} = 1 + \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} \sum_{k=j+1}^{M_i} d_{ij} d_{ik}}{\sum_{i=1}^{N} \sum_{j=1}^{M_i} d_{ij}^2}$$

$$= 1 + \frac{\sum_{i=1}^{N} \left[\bar{d}_i^2 M_i (M_i - 1) - M_i s_{dwi}^2 \right]}{\sum \sum d_{ij}^2} \qquad (4)$$

where s^2_{dwi} is the variance of d within cluster i, i.e.

$$s^2_{dwi} = \frac{1}{M_i} \sum_{j=1}^{M_i} (d_{ij} - \bar{d}_i)^2$$

where $\bar{d}_i$ is the mean value of d within cluster i.

<u>deff</u> can also be formulated using the intracluster correlation of d, ρ_d, as in (1b):

$$\underline{deff} = 1 + \rho_d \left[\bar{M} (1 + v_M^2) - 1 \right] \qquad (5)$$

While ρ_d is more cumbersome to compute than is <u>deffr</u> (using equation (3)), this formulation permits an intuitive understanding of <u>deff</u> Define $R_i = \bar{y}_i / \bar{x}_i$ = ratio for a particular cluster. Of course, due to randomness, no one would expect than all the $R_i = R$. If the cluster elements are independent of each other, the degree to which the R_i's should vary about R can be determined from the variation from observation to observation. If the R_i's vary less than this amount, i.e. if most of the R_i's are very close to R, ρ_d will be negative, implying <u>deff</u> < 1. However, if the R_i's vary a lot, with some clusters having high R_i's and other low R_i's, ρ_d will be positive and <u>deff</u> > 1.

<u>Application of Cluster Sampling Theory</u>

The only obstacle to applying the foregoing theory is estimating the design effect (or the intracluster correlation). (The term v_M^2 can easily be computed by simulating cluster selection, and is often small enough compared to 0 as to be insignificant.) Transit systems that have practiced cluster sampling can estimate <u>deff</u> and <u>deffr</u> from historical data. For systems that lack the data or the ability to estimate these factors, a set of default values would be useful.

The PAT data afforded an opportunity to estimate the design effect of cluster sampling for the case of data collected by run, as ride checks usually are. The data allowed us to calculate the effect on the estimates of means and ratios (but not on category proportions).

We looked at clusters for statistics measured at two levels: at the level of route/direction/time period (R/D/TP), and at the level of route only. In both cases, a cluster consists of a set of trips checked on the same run on the same date for a particular R/D/TP or route.

Sampling by Run for Route/Direction/Time Period Level Statistics

In the case of R/D/TP level data, cluster sampling by run had nearly no effect because cluster size was so small. Weekday time periods are only a few hours in length, and most runs have a very small number of trips in a given direction during a single time period. In the PAT data, 94% of the R/D/TP clusters (n=1283) had only 2 or 3 trips. And because trips of the same run within a R/D/TP tend to cover "peak" and "shoulders" of a time period evenly, there is some basis to expect a beneficial effect of sampling by run. Larger clusters, and consequently a negative clustering effect, were expected for the Saturday and Sunday periods, which last all day.

The PAT data had 69 R/D/TP's with at least 8 clusters apiece. Each cluster consisted of 2 or more trips. Since most of the clusters were very small, mean cluster size was only 2.3 trips. We calculated the design effect for the boardings/maximum load ratio for each R/D/TP. (Maximum load = max load on a trip, not load at a prespecified point.) A statistical summary of the results is presented in Table 2. The overall average deff was 1.0, indicating a neutral effect of sampling by run. As expected, a larger design effect was observed in the all-day weekend periods (average deff = 1.19 for Saturday, 1.10 for Sunday). The greatest deff for a single R/D/TP was

Table 2

Design Effect Due to Sampling by Run for

Route/Direction/Time Period Statistics

Statistic = ratio of boardings/maximum load

mean cluster size = 2.3 (minimum = 2, maximum = 8, C.O.V. = 0.18)

mean number of clusters per R/D/TP = 13.6 (min = 8, max = 33)

Time Period	Number of R/D/TP's	Mean deff	Std dev of deff	Minimum deff	Maximum deff	Mean rho*
early am	4	0.96	0.18	0.78	1.21	-0.03
am peak	5	0.60	0.35	0.40	1.22	-0.40
base	25	1.00	0.33	0.52	1.60	-0.01
pm peak	3	1.29	0.36	1.00	1.69	0.29
evening	7	0.71	0.33	0.16	1.10	-0.22
Saturday	15	1.19	0.43	0.56	1.75	0.11
Sunday	10	1.10	0.29	0.62	1.72	0.06
OVERALL	69	1.01	0.38	0.16	1.75	-0.01

* rho = intracluster correlation

1.75. A further test showed no difference between the average deff for better patronized R/D/TP's (mean load $\geqslant$ 25) and less well patronized R/D/TP's.

Based on these results, it seems safe to say that cluster sampling for R/D/TP level statistics can be considered to be just as good as simple random sampling for weekday time periods. For all-day weekend periods, sample size should be increased by 20%. This additional burden is small enough that it probably is outweighed by the benefits of sampling by run in most cases.

Because of the small cluster sizes and neutral design effect found for the R/D/TP-level boardings/maximum load ratio, no further investigation for R/D/TP-level statistics was done.

Sampling by Run for Route Level Statistics

We also tested the effect of sampling by run for route level statistics. Aside from an interest in route level statistics *per se*, PAT's proposed method for calculating system level Section 15 statistics involves the expansion of route level statistics.

For this analysis, only clusters of 4 or more trips (done on the same route, run, and date) were admitted. There were 18 routes with at least 8 clusters (mean = 40 clusters per route). The mean cluster size was 5.0, as 4- and 5-trip clusters accounted for 75% of all the clusters with at least 4 trips. This average seems to indicate either a practice of sampling not entire runs (which last over 7 hrs and contain about 10 trips) but pieces of runs that last 3-4 hours, or that runs are interlined and therefore each run produces two half-run clusters. The results should therefore be interpreted with caution for systems intending to sample entire non-interlined runs.

Table 3 summarizes the results. Two ratios, boardings/ maximum load and passenger-miles/maximum load, were examined, along with one mean, boardings. The average design effect is

Table 3

Design Effect Due to Sampling by Run for

Route Level Statistics

Mean cluster size = 5.0 (minimum = 4, maximum = 16, C.O.V. = 0.26)

Mean number of clusters per route = 39.8 (min = 10, max = 97)

Number of routes = 18

Statistic	Mean deff	std dev of deff	minimum deff	maximum deff	mean rho*
ratio: boardings/ maximum load	1.24	0.31	0.70	1.84	0.056
ratio: pass-mi/ maximum load	1.10	0.40	0.38	1.80	0.018
mean boardings	1.32	0.67	0.12	3.20	0.065

* rho = intracluster correlation

moderate, calling for sample size increases of 10% to 32%. However, the variation is quite wide, especially for mean boardings, where deff varies from 0.12 for one route to 3.20 for another. The range for the ratios is smaller, with the highest deff calculated to be 1.84.

Our results were extended to cover situations in which the average cluster size differs from 5. The average design effect was calculated using the measures of intracluster correlation (rho) and C.O.V. of cluster size calculated from the PAT data. The resulting figures were then inflated a little to make them somewhat conservative, considering the large amount of variation between routes. The resulting recommended design effects are tabulated in Table 4. As an example, suppose a transit system will sample entire runs, averaging 10 trips per day on each run on a single route. If sample sizes are calculated using formulas that assume simple random sampling, they should be inflated by 60% for the boardings/maximum load ratio, by 25% for the pass.-mi/maximum load ratio, and 75% for mean boardings.

With the necessary increase in sample size, is sampling by run a cost-effective strategy? The cost of simple random sampling is probably twice the cost of sampling by run (compare the cost of sampling 4 randomly chosen trips on a route on randomly chosen days to that of sampling 8 trips on a single run on a single day), suggesting that sampling by run is probably cost-effective notwithstanding the greater sample sizes. Sampling pieces of runs appears to be the most efficient option, as it captures most of the efficiency of sampling by run, but has a smaller design effect than sampling entire runs.

Table 4

Recommended Design Effect for Sampling by Run

for Route Level Statistics

	Recommended design effect for:		
Average number of one-way trips sampled per run	ratio of boardings/ maximum load	ratio of pass-mi/ maximum load	mean boardings
3	1.15	1.10	1.20
5	1.30	1.15	1.40
10	1.60	1.25	1.75
15	2.00	1.35	2.20

3. DEFAULT FORMULA FOR LOAD-BOARDINGS CORRELATION

An important determinant of sample size when using conversion factors is the estimated correlation coefficient r_{XY} between the inferred and auxiliary variables. The Transit Data Collection Design Manual (TDCDM) presents the formula for calculating r_{XY}. However, the data required by this formula, namely a paired set of at least 10 observations of the two variables, may not be available prior to data collection. This section reports on efforts to develop a reliable formula for estimating the correlation coefficient for a particular pair of inferred and auxiliary items, peak load and boardings. (It doesn't matter which one is the auxiliary and which the inferred variable, since $r_{XY} = r_{YX}$.)

First, the intrinsic relationship between peak load and boardings was exploited to yield a theoretical estimate of r_{XY}. Because this estimate, called a_{XY}, is based on an assumption that does not hold perfectly, a systematic discrepancy still remained between a_{XY} and r_{XY}. Polynomial regression was then used to get a better, unbiased fit. The result is a formula for r_{XY} whose standard error is small enough to make it useful in data collection design.

The data used in this analysis were the Pittsburgh ride checks described elsewhere in this report.

Formulas for Theoretical Estimate of r_{XY}

People who are on board the bus at the peak load point (or any other point, for that matter) are a subset of the people who board the bus anywhere on the route. Let

X = load (at the peak point)
Y = boardings
$p = X/Y$ = ratio of mean peak load to mean boardings
v_X, v_Y = C.O.V. of load, of boardings
r_{XY} = correlation coefficient between X and Y.

Each of the Y persons boarding the bus can be considered a coin toss with the two possible outcomes being "on board at the PLP (peak load point)" and "not on board at the PLP" (heads or tails). If each person is assumed to have the same probability p of being on board at the PLP (remember, we assume no knowledge of where people board or alight), and if people are assumed to behave independently (e.g. negligible group travel), then X can be considered a Binomial random variable with parameters p and Y.

Based on the above assumptions, the theoretical correlation coefficient between X and Y can be expressed as a relationship between Y, p, and v_Y:

$$a_{XY} = \frac{v_Y}{\sqrt{\frac{1-p}{p\bar{Y}} + v_Y^2}}$$

(Note that p is the estimated conversion factor if the inferred item is peak load, and that (1/p) is the conversion factor if the inferred item is boardings.) Notice that this equation uses information about Y (boardings) but not X. An alternative theoretical estimate is

$$a'_{XY} = \frac{\sqrt{v_X^2 - \frac{1-p}{\bar{X}}}}{v_X}$$

These formulas are derived in Appendix C. These theoretical estimates are consistent in that a_{XY} and a_{XY}, like r_{XY}, cannot be greater than 1, and equal 1 when p=1 (i.e., when X=Y).

Plotting either a_{XY} or a'_{XY} versus r_{XY} revealed the same trend in both the Fall '83 and Fall '84 datasets. As Figure 2 illustrates, the relationship is strong, but a_{XY} and a'_{XY} systematically overestimate r_{XY}. Therefore, a statistical fit between a_{XY} and r_{XY} was sought. The fit was constrained to pass through the point (a_{XY}=1, r_{XY}=1).

Several functional relationships were explored, including linear, log-linear, exponential, and reciprocal. Only route/direction/time periods with at least 10 observations were allowed. The best fit was found to be the polynomial expression

$$(1-\hat{r}_{XY}) = b_1(1-a_{XY}) + b_2(1-a_{XY})^2$$

which simplifies to

$$\hat{r}_{XY} = c_0 + c_1 a_{XY} + c_2 (a_{XY})^2$$

where $c_0 = 1-b_1-b_2$, $c_1 = b_1+2b_2$, and $c_2 = -b_2$.

The results from the Fall '83 dataset are

c_0	c_1	c_2	std. error	n	mean r_{XY}
0.28	-0.53	1.25	0.046	108	0.941

Similar analysis using a'_{XY} instead of a_{XY} yielded these results:

c_0	c_1	c_2	std. error	n	mean r_{XY}
1.82	-3.64	2.82	0.037	108	0.941

Based on standard error, the load-based approach (i.e., using a'_{XY}) was slightly preferred over the boardings-based approach.

Figure 2 shows the curve relating the load-based estimate against r_{XY}. The plot shows a tendency to overestimate r_{XY} when r_{XY} is low, and to underestimate it when it is high. Because the most serious error we would like to avoid is overestimating r_{XY} when it is low, we restimated r_{XY} using only points where $r_{XY} \leq 0.95$. The results are:

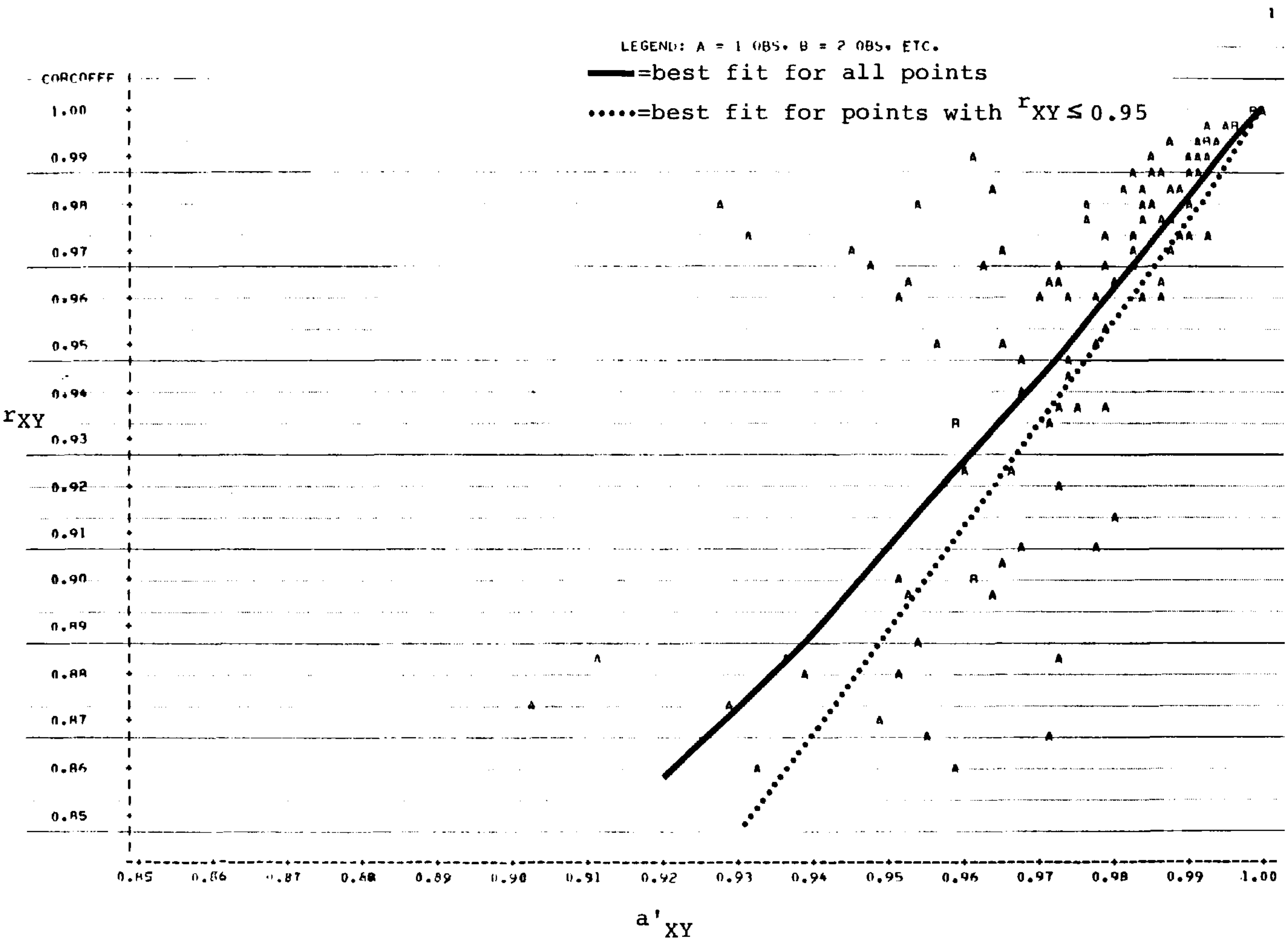
c_0	c_1	c_2	n	std. error for entire dataset (n=108)
2.56	-5.63	4.075	42	0.041

This relationship is also shown in Figure 2.

In summary, the correlation coefficient r_{XY} between peak load (X) and boardings (Y) can be estimated using default formulas when a set of paired data for direct estimation is unavailable. This default r_{XY} can thus be used in the equations for determining the size of the baseline paired sample. After the sample is taken, however, r_{XY} should be estimated directly from the baseline data (using the TDCDM formula) and compared with r_{XY}. If r_{XY} r_{XY}, the sample size may have been larger than necessary; if r_{XY} r_{XY}, it may have been too small. The formulas for r_{XY} are slightly conservative in that they offer a chance of about 30% of the sample being too large and a chance of about 20% of the sample being too small. The transit system that desires to be more conservative in its sample size determination can subtract 0.04 from r_{XY}. Our results suggest that with this adjustment, the sample size will be adequate about 90% of the time, though it may be excessive 40% of the time.

Figure 2

LOAD-BOARDINGS CORRELATION COEFFICIENT (r_{XY}) VERSUS LOAD-BASED THEORETICAL ESTIMATE (a'_{XY})



4. OPTIMAL SIZING OF BASELINE SAMPLE FOR CONVERSON FACTORS: MORE FLEXIBILITY THAN CURRENT TDCDM

In the Transit Data Collection Design Manual (TDCDM), the baseline sample size for estimating conversion factors is given by equations (6.13) and (6.14). These formulas are based on an optimizing framework in which the costs of the baseline phase are balanced against those of the monitoring phase. For simplicity's sake, the optimizing framework is omitted from the manual, and equations (6.13) and (6.14) are the results for what we considered to be a "typical" application. However, Pittsburgh's situation differs substantially from this "typical" prototype, and differing scenarios will undoubtedly appear. Therefore, it seems important to expand on the treatment of this subject so that TDCDM users have more flexibility for applying its procedures to their situation.

The remainder of this section could replace the top half of page 93 of TDCDM, or serve as a supplement.

In achieving the desired tolerance of an inferred item in the monitoring phase, the transit agency controls two variables: n_1, the size of the baseline paired sample (which determines the accuracy of the conversion factor), and n_2, the size of the monitoring sample (which determines the accuracy of the auxiliary item). There is a cost tradeoff, i.e., the greater n_1 is, the smaller n_2 need be. To minimize its overall costs, both should be considered together. Three cases are examined.

<u>Case A: perfect accuracy (complete sample) of auxiliary item</u>. If the auxiliary item will be known with complete accuracy because it is routinely collected on every trip,

then the only factor affecting the accuracy of the inferred item in the monitoring phase is n_1. To ensure that the inferred estimate will have a desired tolerance d_m, the number of paired observations in the baseline sample should be:

$$n_1 = 1.7 + \frac{3.24}{d_m^2}\left(v_X^2 + v_Y^2 - 2 r_{XY} v_X v_Y\right) \qquad (6.14a)$$

<u>Case b: the auxiliary item will be sampled</u>. If the auxiliary item will be sampled at some cost (e.g. through point checks), then the auxiliary item will have less than perfect accuracy, and so the quantity given by (6.14a) is strictly a lower limit on the baseline sample size; that is, a greater sample size will likely be needed. The formulas given below yield the minimum cost plan, accounting for data collection costs in both the baseline and monitoring phases.

Let

c = ratio of the unit cost of a monitoring sample to the unit cost of a baseline sample.

F = number of times during the life of the conversion factor that a monitoring sample will be taken.j

For example, if monitoring is to be done with point checks, with the checker monitoring trips in both directions, and if the service headway is h minutes, then the monitoring sample requires h checker-minutes per pair of trips. And if the baseline sample is to be done with ride checks, and if ride checks are to be done in both directions and round trip running-time (including layover) is T minutes, the unit baseline cost is T checker-minutes per trip pair. Thus, in this case

$$c = \frac{h/2}{T/2} = \frac{h}{T}$$

The variable F can be determined by multiplying the expected life of the conversion factor (in years) by the number of times per year the auxiliary item is estimated. For example, if the conversion factor will be used for 4 years before it is reestimated (with another baseline sample), and if the monitoring phase is repeated twice a year, then F = 8.

The relative overall cost of the data collection program will then be

$$\text{Total Cost} = n_1 + Fcn_2$$

Using the relationship between n_2 and n_1 (given in equations (6.15) and (6.10)), the total cost is minimized when

$$n_1 = 1.7 + \frac{3.24}{d_m^2}\left[(v_X^2 + v_Y^2 - 2r_{XY}v_Xv_Y) + v_X\sqrt{cF}\sqrt{v_X^2 + v_Y^2 - 2r_{XY}v_Xv_Y}\right]$$

To simplify the computation, a dummy variable L may be first computed:

$$L = \frac{3.24}{d_m^2}\left(v_X^2 + v_Y^2 - 2r_{XY}\,v_X\,v_Y\right) \qquad (6.13)$$

Then

$$n_1 = L + 1.7 + \frac{1.8\,v_X}{d_m}\sqrt{LcF} \qquad (6.14b)$$

Case c: a "typical" case: baseline ride checks, monitoring point checks. Following the example cited in Case (b) above, if ride checks are used in the baseline phase (both directions are sampled) and if point checks are used in the monitoring phase (both directions monitored by a single checker), then c = h/T. Applying the fundamental fleet size relationship that the number of buses needed on a route is the cycle time divided by the headway, we get c = 1/B where

B = number of buses operated full-time on the route (during the relevant time period)

Then if the conversion factor will be applied quarterly for 4.5 years, F = 4(4.5) = 18 and equation (6.14b) becomes

$$n_1 = L + 1.7 + 7.6 \frac{v_X}{d_m} \sqrt{\frac{L}{B}} \quad \left(\begin{array}{c}\text{round up to} \\ \text{at least 10}\end{array}\right) \qquad (6.14c)$$

Please note that equation (6.14c) is TDCDM equation (6.14).

5. EVALUATION OF PITTSBURG'S PROPOSED SAMPLING PLAN FOR SECTION 15 PASSENGER-MILES

Pittsburgh's transit authority PAT is fortunate in that its longstanding policy is for drivers to count boardings on every trip, every day. Therefore, there is no need for sampling to estimate systemwide boardings. PAT's proposed approach to estimate passenger-miles (PM) is to make ride checks on a sample of trips, compute an average trip length (ATL) from this sample, and then expand it by total boardings. ATL, being the ratio of mean PM to mean boardings, is a conversion factor as described in the Transit Data Collection Design Manual (TDCDM).

In PAT's sampling plan proposed in 1982, 2 routes are selected each week For each selected route, 2 runs are selected on a single weekday, one run being an early run, the other a late run. Weekends are sampled separately; 1 route is selected every other week, and for the selected route one weekend run is chosen. The annual results is that of PAT's 150 or so routes, 100 are sampled on weekdays, and of the 100 routes that operate on weekends, 25 are sampled.

As it began its ride check program, however, PAT found the ride checks so useful for internal purposes that they nearly doubled their overall sample size. Now every route is sampled annually on weekdays, and the weekend rate is higher than 25% (but still below 100%).

For a sampling strategy to be acceptable to UMTA, it must be unbiased and sufficiently accurate ($\pm$10% tolerance at the 95% confidence level). Bias is dependent on sample selection; accuracy on sample size. These issues are discussed separately below.

Sample Selection Procedure

The PAT proposed method has several departures from simple random sampling. Simple random sampling means that every unsampled trip is selected with equal probability and independent of which other trips are selected. PAT's sample selection procedure has the following features:

1. Trips are sampled by run. This is cluster sampling, and will yield an unbiased estimate if runs are selected at random and if the results are expanded and combined properly. The proper way to expand and combine the results is to sum the PM over all the sampled runs, sum the boardings over all the sampled runs, and then take their ratio. Multiplying this ratio (ATL) by total route boardings then yields total route PM.
2. Swing runs (runs with a long break between an a.m. piece and a p.m. piece) and trippers are omitted. This kind of omission biases the sample. The omitted runs tend to include mainly peak period trips, and so their omission biases the sample in favor of off-peak travel patterns. (If data are collected by time period, so that a separate ATL is calculated for each time period and each time period is expanded separately, the bias largely disappears. However, this is not what PAT has proposed.)
3. Interlined runs are omitted. Their omission probably does not compromise route totals on routes with little interlining, because trips belonging to interlined runs are usually well scattered in a route's schedule. To maintain unbiasedness in a structure of sampling by run, however, a "dummy route" should be created consisting of all interlined runs. (If there are many interlined runs, they should be grouped into several dummy routes, so that each dummy route has the total ridership of an average route.)
4. In the proposed plan for weekdays, and in the proposed and actual plan for weekends, not all routes are sampled. Thus we have two-stage sampling: first, routes are selected, then runs are selected within routes.

5. By selecting one early and one late run, PAT is stratifying the sample. They select randomly within each stratum ("early" and "late"), which is proper, but they do not expand the strata separately, biasing the result in favor of the smaller stratum. Proper expansion entails a separate ATL for each stratum, and expansion to total PM for each stratum by multiplying by the total boardings in the stratum.

   The bias resulting from combining the strata before expanding (i.e. calculating a single ATL for the route) vanishes if the strata are of the same size, i.e. have the same total boardings. PAT could therefore equalize stratum size by moving some runs to the smaller stratum. This might mean, for example, putting a few early runs in the "late" stratum. There is a chance then, for this example, that the run selected from the "late" stratum is in fact an early run.

   In summary, PAT should either equalize the sizes of the two strata, or expand the strata separately.

   Another effect of stratifying runs into an early and late group is to alter the accuracy of the results. Both PAT and we expect that the accuracy increases slightly by this stratification. For lack of data, however, this effect is not further explored.

6. Weekday and weekend runs are separated, again stratifying the sample. Here there is no doubt that weekday and weekend figures should be expanded separately, as they are.

7. When fewer than 100% of the routes are to be covered, routes are selected at random, with this exception: routes not covered the previous year have priority over those covered the previous year. As long as the previous year's selection was random, this approach does not introduce any bias.

8. Routes are selected with equal probability, regardless of their "size", i.e. their total boardings. With this

sampling plan, the routes should be expanded separately before combining. That is, an ATL should be calculated for each route, then multiplied by total route boardings to yield total route PM. Suppose this is done for 25 routes that were sampled, and the results are then to be expanded to the system level involving 100 routes. A system level ATL should be calculated by summing the (estimated) total PM of the 25 sampled routes, and dividing this sum by the total boardings of those 25 routes. This ATL is then multiplied by system boardings to yield system PM.

9. When all the routes are not sampled, what ATL should be used for the unsampled routes? Point 8 above shows one proper approach, which is to find the overall ATL for the set of sampled routes, and expand it to all routes. This approach is unbiased if the selected routes were selected at random.

   PAT suggested a different approach that uses prior information about the routes. They suggested using the ATL of a <u>similar</u> sampled route (similar in length, type, level of usage, etc.) on an unsampled route. This can be considered a stratification into route group (groups of 2 routes), where an ATL is estimated for each stratum. This approach has 2 drawbacks: first, the strata are not defined objectively, and may change from year to year, and from analyst to analyst; and second, the strata are designed to fit the sampling plan rather than vice versa. We agree that the use of prior information can add to the accuracy of an estimate (though we have not been able to estimate how much), and so we suggest how this might properly be done. Routes should be grouped, based on prior information, to maximize within-group homogenaity of ATL. Group size may vary; some groups may have 1 route, others several. If the sampling fraction is f (i.e. a fraction f of all routes are to be sampled), then groups should preferably contain at least 2/f routes (affording a sample of at least 2 routes per group). Then routes should be selected at random within each group, with each group employing the same sampling fraction (as much as possible). An ATL can then be calculated separately for each group and expanded to group total PM, as described in point (8) (with "group" replacing "system"). Group PM totals are then summed to yield the system total.

10. PAT proposed a still different way of inferring ATL for unsampled weekend routes, based on prior information about the relationship between a route's weekday ATL and its weekend ATL. For many routes, past data showed them nearly the same. The approach would then be to get the ratio of the weekend ATL to the weekday ATL from some past year in which both were measured, and then multiply it by the current year's weekday ATL to yield an estimate of the current year's weekend ATL. We have no information on how stable these ATL ratios are over time, and therefore cannot evaluate the accuracy of an estimate so obtained. Until such data is made available and analyzed, PAT should expand weekend routes similarly to but entirely separately from weekday routes, OR it should combine them all into one sampling frame, so that weekends and weekdays are not distinguished. (If weekends are combined with weekdays, weekday runs should be given 5 times the probability of being selected relative to a weekend run.)

Sample Size and Tolerance

Several correct procedures were outlined above. Necessary sample size will be calculated for the principal procedure. This procedure is to do weekday ride checks on a sample of routes, yielding an ATL for each sampled route, which is then expanded to total weekday PM for each sampled route. Those route totals are summed, yielding a total called "total sample weekday PM." This total is divided by total sample weekday boardings, yielding a system-level weekday ATL, which is multiplied by system total weekday boardings to yield system total weekday PM. The same procedure is applied independently for weekends, yielding system total weekend PM. The weekday and weekend subtotals are then summed to yield overall total PM.

11. Calculating the C.O.V.

For a group such as weekend routes or weekday routes, the squared C.O.V. of the group ATL, and hence of the group total PM, is

$$v_{PM}^2 = \frac{1-n/N}{n}\left(v_{B(T)}^2 + v_{PM(T)}^2 - 2\, r_{B,PM(T)}\, v_{B(T)}\, v_{PM(T)}\right) + \frac{1}{n} E\left(v_{Ri}^2\right)$$

where N = number of routes

n = number of routes sampled

$v_{B(T)}$, $v_{PM(T)}$ = between-route C.O.V. of total boardings, total passenger-miles

$r_{B,PM(T)}$ = between route correlation coefficient of total boardings and PM

$E(v_{Ri}^2)$ = average squared C.O.V. of an estimated route-level boardings-to-PM conversion factor

The first term represents the variance contribution from estimating a system-level ATL from a sample of n routes, assuming perfect accuracy on the sampled routes; the second term accounts for the fact that PM is not known perfectly on the sampled routes, but is estimate using route-level ATL's. $E(v_{Ri}^2)$ is best estimated by estimating the route-level ATL for each route (or a sample or routes) using the following formula and then averaging over the routes:

$$v_{Ri}^2 = \frac{deff}{n_i - 1.7}\left(v_{B(t)}^2 + v_{PM(t)}^2 - 2\, r_{B,PM(t)}\, v_{B(t)}\, v_{PM(t)}\right) \quad (5)$$

where the subscript (t) indicates a trip-level statistic, n_1 is the number of paired samples (ride checks), and <u>deff</u> is the design effect from sampling for a ratio by run. For an estimate, $E[v^2_{Ri}]$ can be calculated by using average values of $v_{B(t)}$, $v_{PM(t)}$, and $r_{B,PM(t)}$ in equation (5).

Applying these formulas to PAT with 100 sampled weekday routes (out of 150), we obtain, using the conservative estimates of the statistics shown below,

$N = 150$	$n = 100$
$v_{B(T)} = 1.0$	avg $v_{B(t)} = 0.5$
$v_{PM(T)} = 1.0$	avg $v_{PM(t)} = 0.5$
$r_{B,PM(T)} = 0.95$	avg $r_{B,PM(t)} = 0.90$
<u>deff</u> = 1.5	

n_1 = number of trips in two half-runs = 10

$$E[v^2_{Ri}] \approx \frac{1.5}{10-1.7}\left[.5^2 + .5^2 - 2(.9)(.5)(.5)\right] = 0.0090$$

$$v^2_{PM} = \frac{1-100/150}{100}\left[1^2 + 1^2 - 2(.95)(1)(1)\right] + \frac{.0090}{100} = 0.00076$$

$$v_{PM} = \sqrt{.00076} = 0.027$$

For weekends, the same values are used except N = 100 (operating routes), n = 25 (sampled routes), n_1 = 8 (one full run is sampled), and <u>deff</u> = 1.75, yielding

$$E[v^2_{Ri}] = \frac{1.75}{8-1.7}\left[.5^2 + .5^2 - 2(.9)(.5)(.5)\right] = 0.014$$

$$v_{PM}^2 = \frac{1-25/100}{25}\left[1^2+1^2-2(.95)(1)(1)\right] + \frac{.014}{25} = 0.00356$$

$$v_{PM} = \sqrt{.00356} = 0.059$$

12. Computing tolerance. The tolerance for total PM, at the 95% confidence level, is simply

$$d = 2v_{PM}$$

Therefore, to get a $\pm$10% tolerance, it is necessary that the system-level v_{PM} be no larger than 0.05. However, if system-level PM is obtained by summing subtotals for different groups, the group tolerance may be broader than $\pm$10%, and the group v_{PM} greater than 0.05, as long as their total meets the requirement, as explained below. For the PAT example, we get

weekday: $d = .054$ or $\pm$5.4%
weekend: $d = .118$ or $\pm$11.8%

13. Combining group subtotals. When group subtotals (e.g. a weekday subtotal and a weekend subtotal) are summed to yield system-level total PM, TDCDM equation (5.2) can be applied:

$$d_{sys} = \frac{1}{\Sigma T_i}\sqrt{\Sigma\, d_i^2 T_i^2}$$

where d_{sys} = system-level tolerance, d_i = tolerance of group i, and T_i = total PM for group i. T_i can be measured on a relative scale; for example, if group 1 accounts for 80% of the system total and group 2 for 20%, use $T_1 = 0.8$ and $T_2 = 0.2$.

Continuing the PAT application, we combine PAT's weekday and weekend subtotals with the assumption that weekday service accounts for 85% of the system total PM, obtaining

$$d_{sys} = \frac{1}{.85 + .15} \sqrt{(.054^2)(.85^2) + (.118)^2(.15)^2} = 0.049$$

So the system-level tolerance is ±4.9%, far better than the required ±10%. In fact, the number of weekday routes sampled annually could be reduced to 26 (one route every other week, same as the weekend sampling plan) and the system-level tolerance achieved would be ±10%.

Summary

Our analysis shows that PAT's sample size is more than adequate. However, some of its sample selectin procedures introduce bias and some of its expansion procedures make it impossible to assess the accuracy. If PAT does not want to adopt the recommended procedures for its entire ride check program, it should at least use the recommended procedures in selecting a sample large enough to meet UMTA's requirements, and then add whatever further ride checks they want for internal purposes.

Appendix A

Proof for the Estimator s_δ^2

$$s_\delta^2 = \frac{1}{r} \sum_{i=1}^{r} \left[(R_{1i} - R_{0i})^2 - s_{R1i}^2 - s_{R0i}^2 \right]$$

Assume R_{1i} independent of R_{0i} (because they are taken for different samples). Then take expected values:

$$E[s_\delta^2] = \frac{1}{r} \sum_{i=1}^{r} \left\{ E\left[(R_{1i} - R_{0i})^2\right] - E\left[s_{R1i}^2\right] - E\left[s_{R0i}^2\right] \right\}$$

$$= \frac{1}{r} \sum_{i=1}^{r} \left\{ \left[E(R_{1i} - R_{0i}) \right]^2 + Var\left[R_{1i} - R_{0i}\right] - \sigma_{R1i}^2 - \sigma_{R0i}^2 \right\}$$

$$= \frac{1}{r} \sum_{i=1}^{r} \left\{ \delta_i^2 + (\sigma_{R1i}^2 + \sigma_{R0i}^2) - \sigma_{R1i}^2 - \sigma_{R0i}^2 \right\}$$

$$= \frac{1}{r} \sum_{i=1}^{r} \delta_i^2 = Var[\delta]$$

Appendix B

Least Cost Baseline Sample Size and Conversion Factor Life with Adjustment for Increasing Uncertainty

a.

Desired tolerance = ±30%

C_{XY} = 0.005

	no adjustment	β^2 = 0.00064		β^2 = 0.0024	
cv_X^2	n_1	n_1	# years	n_1	# years
0.002	10	10	5	10	5
0.006	10	10	5	10	5
0.010	10	10	5	10	5
0.020	10	10	5	10	5
0.030	10	10	5	10	5
0.040	10	10	5	10	5
0.060	10	10	5	10	5
0.080	10	10	5	10	5
0.120	10	10	5	10	5
0.200	10	10	5	10	5

b.

Desired tolerance = ±30%

C_{XY} = 0.015

	no adjustment	β^2 = 0.00064		β^2 = 0.0024	
cv_X^2	n_1	n_1	# years	n_1	# years
0.002	10	10	5	10	5
0.006	10	10	5	10	5
0.010	10	10	5	10	5
0.020	10	10	5	10	5
0.030	10	10	5	10	5
0.040	10	10	5	10	5
0.060	10	10	5	10	5
0.080	10	10	5	10	5
0.120	10	10	5	10	5
0.200	10	10	5	10	5

Notes: v_X, v_Y = C.O.V. of auxiliary item, of inferred item
r_{XY} = correlation coefficient
$C_{XY} = v_X^2 + v_Y^2 - 2r_{XY}v_Xv_Y$
= annual increment in conversion factor C.O.V.
c = (unit monitoring cost)/(unit baseline cost)

c. Desired tolerance = ± 30%

C_{XY} = 0.03

	no adjustment	β^2 = 0.00064		β^2 = 0.0024	
cv_X^2	n_1	n_1	# years	n_1	# years
0.002	10	10	5	10	5
0.006	10	10	5	10	5
0.010	10	10	5	10	5
0.020	10	10	5	10	5
0.030	10	10	5	10	5
0.040	10	10	5	10	5
0.060	10	10	5	10	5
0.080	10	10	5	10	5
0.120	10	10	5	10	5
0.200	10	10	5	10	5

d. Desired tolerance = ± 30%

C_{XY} = 0.05

	no adjustment	β^2 = 0.00064		β^2 = 0.0024	
cv_X^2	n_1	n_1	# years	n_1	# years
0.002	10	10	5	10	5
0.006	10	10	5	10	5
0.010	10	10	5	10	5
0.020	10	10	5	10	5
0.030	10	10	5	10	5
0.040	10	10	5	10	5
0.060	10	10	5	10	5
0.080	10	10	5	10	5
0.120	10	10	5	10	5
0.200	10	10	5	10	5

e. Desired tolerance = ± 30%

C_{XY} = 0.1

	no adjustment	β^2 = 0.00064		β^2 = 0.0024	
cv_X^2	n_1	n_1	# years	n_1	# years
0.002	10	10	5	10	5
0.006	10	10	5	10	5
0.010	10	10	5	10	5
0.020	10	10	5	10	5
0.030	10	10	5	10	4
0.040	10	10	5	10	4
0.060	10	10	5	15	5
0.080	10	10	5	15	5
0.120	10	10	5	15	5
0.200	10	10	5	15	5

f.

Desired tolerance = ± 30 %

C_{XY} = 0.2

	no adjustment	β^2 = 0.00064		β^2 = 0.0024	
cv_X^2	n_1	n_1	# years	n_1	# years
0.002	10	10	5	15	5
0.006	10	15	5	15	4
0.010	10	15	5	15	4
0.020	10	15	5	20	5
0.030	15	15	5	20	5
0.040	15	15	5	20	5
0.060	15	15	5	20	5
0.080	15	15	5	25	5
0.120	15	15	5	25	5
0.200	15	20	5	25	5

g.

Desired tolerance = ± 30 %

C_{XY} = 0.3

	no adjustment	β^2 = 0.00064		β^2 = 0.0024	
cv_X^2	n_1	n_1	# years	n_1	# years
0.002	15	15	5	20	4
0.006	15	15	5	25	5
0.010	15	15	5	25	5
0.020	15	20	5	25	5
0.030	15	20	5	25	5
0.040	15	20	5	30	5
0.060	20	20	5	30	5
0.080	20	20	5	30	5
0.120	20	20	5	30	5
0.200	20	25	5	40	5

h.

Desired tolerance = ± 30 %

C_{XY} = 0.5

	no adjustment	β^2 = 0.00064		β^2 = 0.0024	
cv_X^2	n_1	n_1	# years	n_1	# years
0.002	20	25	5	30	4
0.006	25	25	5	40	5
0.010	25	25	5	40	5
0.020	25	25	5	40	5
0.030	25	30	5	40	5
0.040	25	30	5	40	5
0.060	25	30	5	50	5
0.080	25	30	5	50	5
0.120	30	30	5	50	5
0.200	30	40	5	50	5

i.

Desired tolerance = ±20 %

C_{XY} = 0.005

cv_X^2	no adjustment n_1	β^2 = 0.00064 n_1	β^2 = 0.00064 # years	β^2 = 0.0024 n_1	β^2 = 0.0024 # years
0.002	10	10	5	10	4
0.006	10	10	5	10	4
0.010	10	10	5	10	4
0.020	10	10	5	10	3
0.030	10	10	5	10	3
0.040	10	10	5	10	3
0.060	10	10	5	10	3
0.080	10	10	5	10	3
0.120	10	10	5	10	3
0.200	10	10	5	10	2

j.

Desired tolerance = ±20 %

C_{XY} = 0.015

cv_X^2	no adjustment n_1	β^2 = 0.00064 n_1	β^2 = 0.00064 # years	β^2 = 0.0024 n_1	β^2 = 0.0024 # years
0.002	10	10	5	10	4
0.006	10	10	5	10	3
0.010	10	10	5	10	3
0.020	10	10	5	10	3
0.030	10	10	5	10	3
0.040	10	10	5	10	3
0.060	10	10	5	10	2
0.080	10	10	5	10	2
0.120	10	10	5	10	2
0.200	10	10	5	10	2

k.

Desired tolerance = ±20 %

C_{XY} = 0.03

cv_X^2	no adjustment n_1	β^2 = 0.00064 n_1	β^2 = 0.00064 # years	β^2 = 0.0024 n_1	β^2 = 0.0024 # years
0.002	10	10	5	10	3
0.006	10	10	5	10	3
0.010	10	10	5	10	3
0.020	10	10	5	10	2
0.030	10	10	5	10	2
0.040	10	10	5	10	2
0.060	10	10	5	10	2
0.080	10	10	5	15	2
0.120	10	10	5	15	2
0.200	10	15	5	15	2

l.

Desired tolerance = ±20 %

C_{XY} = 0.05

cv_X^2	no adjustment n_1	$\beta^2 = 0.00064$ n_1	# years	$\beta^2 = 0.0024$ n_1	# years
0.002	10	10	5	10	2
0.006	10	10	5	10	2
0.010	10	10	5	10	2
0.020	10	10	5	15	2
0.030	10	10	5	15	2
0.040	10	15	5	15	2
0.060	10	15	5	15	2
0.080	10	15	5	15	2
0.120	10	15	5	20	2
0.200	15	20	5	20	2

m.

Desired tolerance = ±20 %

C_{XY} = 0.1

cv_X^2	no adjustment n_1	$\beta^2 = 0.00064$ n_1	# years	$\beta^2 = 0.0024$ n_1	# years
0.002	15	15	5	25	3
0.006	15	15	5	20	2
0.010	15	15	5	20	2
0.020	15	20	5	20	2
0.030	15	20	5	25	2
0.040	15	20	5	25	2
0.060	15	20	5	25	2
0.080	15	25	5	25	2
0.120	20	25	5	30	2
0.200	20	30	5	30	2

n.

Desired tolerance = ±20 %

C_{XY} = 0.2

cv_X^2	no adjustment n_1	$\beta^2 = 0.00064$ n_1	# years	$\beta^2 = 0.0024$ n_1	# years
0.002	20	25	5	30	2
0.006	20	30	5	30	2
0.010	20	30	5	40	2
0.020	25	30	5	40	2
0.030	25	30	5	40	2
0.040	25	30	5	40	2
0.060	25	40	5	40	2
0.080	30	40	5	50	2
0.120	30	40	5	50	2
0.200	30	50	5	50	2

a.

Desired tolerance = ±20 %

C_{XY} = 0.3

cv_X^2	no adjustment n_1	$\beta^2 = 0.00064$ n_1	# years	$\beta^2 = 0.0024$ n_1	# years
0.002	30	40	5	50	2
0.006	30	40	5	50	2
0.010	30	40	5	50	2
0.020	30	40	5	50	2
0.030	40	50	5	50	2
0.040	40	50	5	50	2
0.060	40	50	5	75	2
0.080	40	50	5	75	2
0.120	40	50	5	75	2
0.200	50	75	5	75	2

b.

Desired tolerance = ±20 %

C_{XY} = 0.5

cv_X^2	no adjustment n_1	$\beta^2 = 0.00064$ n_1	# years	$\beta^2 = 0.0024$ n_1	# years
0.002	50	75	5	75	2
0.006	50	75	5	75	2
0.010	50	75	5	75	2
0.020	50	75	5	75	2
0.030	50	75	5	100	2
0.040	50	75	5	100	2
0.060	50	75	5	100	2
0.080	75	75	5	100	2
0.120	75	75	5	100	2
0.200	75	100	5	100	2

g.

Desired tolerance = ±10 %

C_{XY} = 0.005

cv_X^2	no adjustment n_1	$\beta^2 = 0.00064$ n_1	# years	$\beta^2 = 0.0024$ n_1	# years
0.002	10	10	3	15	1
0.006	10	10	2	15	1
0.010	10	10	2	20	1
0.020	10	10	2	25	1
0.030	10	10	2	25	1
0.040	10	15	2	30	1
0.060	10	15	2	30	1
0.080	10	15	2	40	1
0.120	10	20	2	40	1
0.200	15	20	2	50	1

r.

Desired tolerance = ±10 %

C_{XY} = 0.015

cv_X^2	no adjustment n_1	β^2 = 0.00064 n_1	β^2 = 0.00064 # years	β^2 = 0.0024 n_1	β^2 = 0.0024 # years
0.002	10	15	2	30	1
0.006	10	15	2	40	1
0.010	10	15	2	40	1
0.020	15	20	2	50	1
0.030	15	20	2	50	1
0.040	15	25	2	50	1
0.060	15	25	2	75	1
0.080	20	30	2	75	1
0.120	20	30	2	75	1
0.200	25	40	2	100	1

s.

Desired tolerance = ±10 %

C_{XY} = 0.03

cv_X^2	no adjustment n_1	β^2 = 0.00064 n_1	β^2 = 0.00064 # years	β^2 = 0.0024 n_1	β^2 = 0.0024 # years
0.002	15	25	2	50	1
0.006	15	25	2	75	1
0.010	20	30	2	75	1
0.020	20	30	2	75	1
0.030	20	40	2	100	1
0.040	25	40	2	100	1
0.060	25	40	2	100	1
0.080	25	50	2	100	1
0.120	30	50	2	100	1
0.200	40	75	2	100	1

t.

Desired tolerance = ±10 %

C_{XY} = 0.05

cv_X^2	no adjustment n_1	β^2 = 0.00064 n_1	β^2 = 0.00064 # years	β^2 = 0.0024 n_1	β^2 = 0.0024 # years
0.002	20	40	2	100	1
0.006	25	40	2	100	1
0.010	25	40	2	100	1
0.020	30	50	2	100	1
0.030	30	50	2	100	1
0.040	30	50	2	100	1
0.060	40	50	2	100	1
0.080	40	75	2	100	1
0.120	40	75	2	100	1
0.200	50	75	2	100	1

u.

Desired tolerance = ± 10 %

C_{XY} = 0.1

	no adjustment	$\beta^2 = 0.00064$		$\beta^2 = 0.0024$	
cv_X^2	n_1	n_1	# years	n_1	# years
0.002	40	75	2	conversion not feasible	
0.006	40	75	2		
0.010	50	75	2		
0.020	50	75	2		
0.030	50	100	2		
0.040	50	100	2		
0.060	50	100	2		
0.080	75	100	2		
0.120	75	100	2		
0.200	75	100	2		

v.

Desired tolerance = ± 10 %

C_{XY} = 0.2

	no adjustment	$\beta^2 = 0.00064$		$\beta^2 = 0.0024$	
cv_X^2	n_1	n_1	# years	n_1	# years
0.002	75	100	1	conversion not feasible	
0.006	75	100	1		
0.010	75	100	1		
0.020	100	100	1		
0.030	100	100	1		
0.040	100	100	1		
0.060	100	100	1		
0.080	100	100	1		
0.120	100	100	1		
0.200	100	100	1		

w.

Desired tolerance = ± 10 %

C_{XY} = 0.3

	no adjustment	$\beta^2 = 0.00064$		$\beta^2 = 0.0024$	
cv_X^2	n_1	n_1	# years	n_1	# years
0.002	100	conversion not feasible		conversion not feasible	
0.006	100				
0.010	100				
0.020	100				
0.030	100				
0.040	100				
0.060	100				
0.080	100				
0.120	100				
0.200	100				

Appendix C

Derivation of Theoretical Correlation Coefficient Estimates

The relationship between boardings and peak load is such that it is possible to consider each boarding passenger a Bernoulli trial with the possible outcomes "on board at the peak point" (success) or "not on board at the peak point" (failure). If passenger behavior in this regard is independent across passengers and consistent across trips (assumptions which are not perfectly net, which is why the derived quantities are estimates), then peak load can be modeled as a binomial random variable, leading to the following deviations.

Let

Y = boardings on a single trip

X = peak load on a single trip

$\bar{Y}$ = mean boardings (per trip)

$\bar{X}$ = mean peak load

p = X/Y

ρ_{xy} = correlation coefficient between X and Y

Under our assumptions, $X_i \sim$ Binomial (Y_i, p). Thence it follows that, for a given trip with Y_i boardings,

$$E[X_i | Y_i] = pY_i$$

$$Var[X_i | Y_i] = p(1-p)Y_i$$

$$E[X_i^2 | Y_i] = p(1-p)Y_i + p^2 Y_i^2$$

$$E[X_i Y_i | Y_i] = Y_i E[X_i | Y_i] = pY_i^2$$

Averaging now over all trips,

$$E[X] = \int_{all\ Y} (pY) f_Y(y)\, dy = pE[Y] = p\bar{Y}$$

$$E[X^2] = \int_{all\ Y} p(1-p) Y f_Y(y)\, dy + \int_{all\ Y} p^2 Y^2 f_Y(y)\, dy$$

$$= p(1-p)\bar{Y} + p^2 E[Y^2]$$

$$= p(1-p)\bar{Y} + p^2 Y^2 + p^2 \sigma_Y^2$$

$$Var[X] = E[X^2] - (E[X])^2 = p(1-p)\bar{Y} + p^2 \sigma_Y^2 \qquad (C.1)$$

$$E[XY] = \int_{all\ Y} pY^2 f_Y(y)\, dy = pE[Y^2] = p(\bar{Y}^2 + \sigma_Y^2)$$

$$Cov[XY] = E[XY] - E[X]E[Y] = p\bar{Y}^2 + p\sigma_Y^2 - p\bar{Y}^2 = p\sigma_Y^2$$

$$\rho_{XY} = \frac{Cov[XY]}{\sigma_X \sigma_Y} = \frac{p\sigma_Y^2}{\sigma_X \sigma_Y} = p\frac{\sigma_Y}{\sigma_X}$$

Making the substitutions $\sigma_Y = v_Y \bar{Y}$ and $\sigma_X = v_X \bar{X}$,

$$\rho_{XY} = \frac{p v_Y \bar{Y}}{v_X \bar{X}} = \frac{p v_Y \bar{Y}}{v_X p \bar{Y}}$$

$$\rho_{XY} = \frac{v_Y}{v_X} \qquad (C.2)$$

This formula for estimating ρ_{xy} relies on the C.O.V.'s of both X and Y, which are likely to be rough estimates themselves. Since transit systems often have far better data on one variable (i.e. X or Y) than on the other, two variations of equation (C.2) were derived, using the following two substitutions. The first substitution follows from equation (C.1); the second follows from the first. The substitutions are:

$$v_X^2 = \frac{p(1-p)\bar{Y} + p^2\sigma_Y^2}{p^2\bar{Y}^2} = \frac{1-p}{p\bar{Y}} + v_Y^2$$

$$v_Y^2 = v_X^2 - \frac{1-f}{f\bar{Y}} = v_X^2 - \frac{1-f}{\bar{X}}$$

(Notice the implication that the C.O.V. of peak load should exceed the C.O.V. of boardings.) The resulting formulas for ρ_{xy} are:

$$\rho_{XY} = \frac{v_Y}{\sqrt{\frac{1-f}{f\bar{Y}} + v_Y^2}}$$

$$\rho_{XY} = \frac{\sqrt{v_X^2 + \frac{1-f}{\bar{X}}}}{v_X}$$