Enhancing Patronage Estimation and Line Performance Monitoring Procedures

Technical Memorandum 1

Prepared for:

Southern California Rapid Transit District

September 1987

Multisystems, Inc.

1050 Massachusetts Avenue

Cambridge, Massachusetts 02138

1. Description and Evaluation of Current Procedures

1.1 Description of Current Sampling Plan

SCRTD's current sampling plan and estimation procedures for estimating systemwide patronage relies on quarterly fare checks. In the selected month (usually Feb, May, Aug, and Nov), a "typical" week is selected, and the fare check is conducted during this week.

The sampling unit is half a run (run = a vehicle tour). Lines are divided into four strata. Three are used for local bus lines and fourth stratum is used for express lines. The local bus strata are determined by the average number of boardings per bus trip. Stratum 1 consists of bus lines averaging more than 100 boardings per trip. Stratum 2 consists of lines averaging 50 to 100 boardings per trip and stratum 3 consists of lines averaging less than 50 boardings per trip. Weekdays, Saturdays and Sundays are sampled and expanded separately.

Until recently, equal numbers of half-runs were selected in each stratum. The sample size was chosen to be quite large so that there would be little question of its adequacy; however, there was no scientific basis to determine how much the sample size could be reduced. Sample selection was done by randomly selecting a trip from the trip file, and then sampling the half-run beginning with this trip. This procedure makes the probability of a half-run being selected proportional to the number of trips in the run. If the selected trip fell in the second half of the run, a half-run beginning with that trip would have to wrap-around to the following morning. Because of the high cost of checking wrap-around half-runs, the selection procedure was modified by pushing the start of the half-run back to the midpoint of the run when the selected trip fell in the second half of a run. Unfortunately, this adustment leads to an oversampling of mid-day and evening peak trips, and so a new selection procedure was introduced in the Spring, 1987 fare

check. In the new procedure, runs are presplit into an early and a late half, and half-runs are selected with equal probability.

For weekdays the selected half-runs are distributed evenly, but randomly, across the five weekdays of the fare check week. Boardings are recorded, classified into some 123 categories, each of which has an associated unique fare.

By multiplying boardings by the appropriate fare in each category, revenue is estimated for each trip. We shall call revenue thus estimated "assumed revenue", since it may differ from actual revenue due to under- and over-payments. (New fare check procedures call for recording under- and over-payments, but due to observation errors, assumed revenue will still not be identical to actual revenue.)

In expanding the results, the 123 categories are compressed to 4 pass categories and a non-pass category. Expansion for the non-pass category is done as follows. Non-pass boardings and revenues are expanded at the stratum level (dividing by the fraction of trips sampled) and summed to yield system daily non-pass boardings and assumed revenue. Their ratio is AVNPB = average value of a non-pass boarding. Expanding using total actual revenue for the month yields the monthly estimate of non-pass boardings.

The pass category boardings are similarly expanded at the stratum level and summed, yielding daily pass category use. Dividing by pass sales during the fare survey month yields the number daily uses per pass in four pass categories, which is multipled by number of days of the appropriate day type (weekday, Sat., Sun.) in the month to yield monthly use.

In months between fare surveys, the number of daily uses per pass and AVNPB are linearly interpolated between the straddling fare survey months.

When making a daily estimate of patronage, a day's actual revenue is multiplied by the month's AVNPB value to yield

non-pass boardings. Pass boardings is the product of number of passes in circulation and average daily uses per pass. After the 10th of the month, number of passes in circulation is assumed equal to total pass sales for the month. However, before the 10th of the month, all passes are not yet in circulation. A small sample of RTD outlets provided the planning staff with data indicating pass sales by day, from which a curve showing fraction of passes in circulation by date has been derived. For example, only a little over 50% of the passes sold in a month are in circulation on the first of the month. This curve is used to estimate number of passes in circulation from total sales during the first 10 days of the month.

Auxiliary data used, i.e., pass sales, daily revenue, and expansion factors relating to numbers of trips, all appear to be reliable.

1.2 Evaluation of Current Sampling and Expansion Plan

The following comments and recommendations are offered regarding current procedures.

a. Until recently, trippers and interlined runs were excluded from the sample. To guarantee unbiasness, they have been added to the sampling frame beginning in Spring, 1987. Choosing the stratum of an interlined run can be done in any reasonable, consistent way. By consistent, we simply mean that the interlined runs be classified before sample selection, and that expansion recognize the presence of these runs in the strata into which they were classified. For example, the run could be classified with the parent line, or according to an average of boardings on the parent and foreign lines. When an interlined run consists of an early piece on one line and a late piece on another, the run could be split into unequally sized half-runs in the half-run sampling frame; alternatively, as long as the number of interlined runs is small, they could be split after sampling (as in the ex post facto strtaification approach described later) and intercluster correlation effects can be ignored.

b. Runs should be presplit into an early and a late half. The halfs need not contain equal numbers of trips; the runs may be split at convenient points. Trippers or other short runs may be kept whole, and especially long runs may be cut into 3 or more pieces. This way clusters (i.e., pieces) can be directly selected according to the sampling strategy chosen, i.e., either with equal probability or with probability proportional to cluster size (pps).

c. The two sample selection strategies, sampling clusters with equal probability and pps sampling, have different expansion techniques, and it is important to use the correct technique (described in Appendix A). Until recently, SCRTD used the pps expansion procedure with a pps sample, and is now doing equal probability sampling with the equal probability expansion procedure. There was, however, a six-month transition time in which the equal probability sampling procedure was used with pps sampling. This error has since been corrected, and should continue to be guarded against.

d. Doing the fare survey in a "typical" week biases the results because a typical week is not an average week, i.e., does not contain an average number of minor holidays, school vacation days, busy shopping days, days in the start of the month when pass sales are only partial, etc. Instead, the fare survey should be spread over the whole month, either on every day or on randomly selected days.

e. Interpolating between months makes it difficult to assess accuracy of estimates based on interpolation. It may be that average daily use per pass is higher in December than in other months, and this can not be measured with a quarterly survey. Also, the 4 months selected are not "average" months. For these reasons, it is better to spread the fare survey out over 12 months of the year, making fresh estimates in each month. While this change will ostensibly triple the sample size to meet a monthly precision level, it will greatly increase the accuracy of the annual estimate and provide reliable estimates in 12 months, rather than in only four. Alternatively, sample size could be left unchanged if what is desired is a good estimate of quarterly boardings.

f. Using actual revenue to expand a factor based on assumed revenue has two effects. First, it systematically biases the estimates to the degree assumed revenue is systematically higher or lower than actual. Second, it lowers the precision to the extent that assumed revenue varies randomly about actual revenue. There is a feeling, and there is some evidence from limited SCRTD studies, that there is a good correlation between actual and assumed revenue. This correlation should be better confirmed, however.

g. The current procedures put a labor crunch on checkers four weeks a year. This could be alleviated by spreading the fare survey over the entire year.

h. In expanding estimates using the ratio of (total population trips)/(sampled trips), the number of trips in the population should reflect missed trips by multiplying number of scheduled trips by percentage missed trips.

2. Alternative Sampling and Estimation Approaches

Five approaches were considered for estimating monthly partronage. All are based on keeping weekdays, Saturdays, and Sundays separate, and on sampling continually over the entire year.

2.1 Stratified Cluster Sampling with Sample Total Expansion (Ratio-to-Cluster Size)

This approach is similar to the one now being used by RTD, except that clusters are selected with equal probability. The sample total (within each stratum) is expanded by dividing by the fraction of trips in the stratum that are sampled. This is known as the "ratio to cluster size" approach.

2.2 Stratified Cluster Sampling with Selection Probability Proportional to Size

Half-runs (cluster) are selected within each stratum with probability proportional to the number of trips in the half-run. In each cluster sampled, the average boardings per trip is calculated, and the average of these averages is expanded by the number of trips in the population.

2.3 Unstratified Cluster Sampling Using a Revenue-Based Conversion

From a sample of clusters (selected with equal probability), boardings and revenue are totaled. Their ratio becomes a conversion factor which, when applied to monthly revenue, yields in a monthly boardings estimate.

2.4 Cluster Sampling with Ex-Post-Facto Line/Direction/Time Period Stratification

A shortcoming of the current stratification approach is that even when lines are stratified by boardings, there is still a great deal of variation that can't be eliminated because of how much boardings per trip varies by time period and direction. Prior stratification by line/direction/time period (L/D/TP) is impractical, since half-runs (which always span two directions, and often span two periods) are a natural sampling unit.

Sampling is done by half-run, as before; now, however, a half-run is called a supercluster. Superclusters are selected without stratification, with equal selection probabilities.

Each trip in the system is labeled with the average boardings per trip of the line/direction/time period (L/D/TP) to which it belongs. These L/D/TP averages come from recent ride checks. Stratification is then done by values of these labels. A cluster is now defined to be the group of trips in a supercluster that lie in the same stratum. A cluster may therefore contain trips of more than one line, direction, or period, if the trips belong to the same supercluster and their labels fall in the same stratum.

Expansion is done by getting the sample total in each stratum and dividing by the fraction of trips in the stratum that were sampled.

2.5 Direct Stratification of Clusters

Another way to stratify more finely than by line is to directly stratify clusters (half-runs). Using the most recently available ride check data, each trip in a cluster can be assigned a value of expected boardings. Because of how schedules change, a perfect correspondence of trips in the current schedule to trips in the ride check database is not possible. Therefore, average boardings per trip for each line, direction, and one-hour time period was calculated from the ride check database. This value was then assigned as the expected boardings on trips in that L/D/TP in the current schedule. The expected boardings per trip for a cluster in the current schedule is calculated by averaging the assigned values of the trips in the cluster. Clusters are then stratified by their average expected boardings per trip.

Clusters are then selected randomly with equal probability within each stratum, and expanded using the ratio-to-cluster-size approach, yielding estimates of total boardings for the set of clusters in each of the strata, which are then summed to yield a system estimate.

2.6 Further Stratification

For each of the four methods employing stratification, a possible way of gaining precision is to have more strata.

2.7 Technical Descriptions, Variance and Sample Size Formulas

These are found in Appendix A.

2.8 Preliminary Evaluation of the Approaches

Before analyzing the data, we had the following expectations concerning the different approaches. The first two approaches, both of which involve cluster sampling with line stratification, were expected to be equally efficient.

The revenue-based approach was expected to be superior on the basis that knowing the revenue of a trip is more valuable than knowing the stratum of the line the trip belongs to in predicting boardings. However, the problem of estimating the revenue conversion factor with assumed revenue and expanding with actual revenue introduces some additional variance, the extent of which was not known. Line/direction/time period stratification is also expected to perform better than line stratification. Further stratification also was expected to improve precision to a point; however, this approach is limited in that very small sample sizes within a stratum (fewer than four clusters) were deemed impractical.

3. Sampling of Sampling and Estimation Approaches

Statistical inputs to the variance formulas such as per cluster coefficients of variation, correlation coefficients, and population descriptors, were generated from past fare survey data on SCRTD's mainframe using SPSS. These inputs were then manipulated in LOTUS speadsheets to yield reports on sample size needed, tolerance achieved, and so forth.

The SPSS and LOTUS programs will be described in a separate technical memorandum so that SCRTD staff may do further analyses if desired.

The 95% confidence level is used throughout. If the standard deviation is known, the corresponding z-value is 1.96 (i.e., there is a 95% chance that a standard normal variate lies between -1.96 and +1.96). Because we are using the standard deviations that are estimated from the samples, we have used a z-value of 2.1 throughout. (This may be somewhat conservative, since 2.1 is the t-value for the 95% confidence level with 18 degrees of freedom, whereas the fare survey results supplied between 15 and 200 degrees of freedom, depending on the level of stratification, in estimating standard deviations.)

3.1 Ratio-Cluster-Size Sampling with Stratification by Line

We analysed the November, 1986 and the February, 1987 fare check datasets, weekdays only. At first, we used the same four strata now used by RTD. In general, the February dataset showed more variation, and so, to be conservative, we make our recommendation based on February results. Table 1 shows the precision levels for daily mean system boardings that would have been attained by the November and February samples if half-runs were selected with equal probability, and with the expansion procedure actually used. The November sample, with 198 half-runs sampled overall, yields a precision of $\pm$5.9%, while the February sample, with 194 half-runs sampled and with greater variation within strata, yields a precision of $\pm$7.3%.

Table 2 shows the sample size necessary to attain various levels of precision using the ratio-to-cluster-size approach. Sample size has been optimally allocated among the strata to minimize the number of half-runs sampled overall to achieve a given precision. Input statistics come from the February fare check. Overall 80 half-runs must be sampled to achieve a $\pm$10% precision. The allocation among strata is not at all even, with stratum 1 getting 36 half-runs and stratum 3 getting only 4 half-runs. However, the precision level is quite robust with respect to departures from optimal allocation. For example, if 80 half-runs are split equally among the four strata (20 in each stratum), precision widens only to $\pm$11.6%; and if the 80 half-runs are split among the strata with equal sampling rates (with strata 1, 2, 3, 4 getting 34, 21, 5, and 21 half-runs, respectively), the precision is $\pm$10.1%.

Two notes of caution should be made about these results. First, the analysis was based on a single week of data, and we are using it as though it were a random sample from a month. If weeks tend to be homogeneous in themselves, but quite different between each other, our results are optimistic. If the variation between weeks is small compared to the variation within a week, the results are acceptable. We see no reason to

TABLE 1: Stratified Cluster Sampling
Ratio-to-Cluster-Size Approach
Precision Achieved with Sampled Data

November 1986 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Clusters
1	100	7574	929	4.1	52	117.9	0.26	52
2	50	5531	586	4.7	54	74.6	0.34	54
3	0	1584	125	6.3	44	30.0	0.66	44
4	express	2800	508	2.8	48	47.6	0.93	48

Total Number of Clusters:	198
Expected Number of Trips:	806
COV of Total Boardings:	0.028
95% Precision:	5.9%

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Clusters
1	100	7507	937	4.0	49	111.8	0.32	49
2	50	5535	589	4.7	53	68.0	0.45	53
3	0	1504	126	6.0	44	29.8	0.70	44
4	express	2823	580	2.4	48	48.4	0.94	48

Total Number of Clusters:	194
Expected Number of Trips:	770
COV of Total Boardings:	0.035
95% Precision:	7.3%

expect large week-to-week variation (as compared with day-to-day variation), and therefore feel comfortable in applying the results to a monthly time frame. Second, the data in the analysis datasets depart from being a random sample in two ways: trippers and interlined runs are omitted, and half-runs were not selected with equal probailities. Again, we believe the effects are small. Trippers and interlined runs constitute a small fraction of the daily schedule. The sample selection process used to generate the datasets oversampled runs with many trips, and possibly oversampled the later half of the day. If anything, these departures from simple random sampling should slightly increase the estimated variance, and so its seems safe to accept the results.

Table 3 shows the precision attained for various overall sample sizes, assuming optimal allocation among strata. This report shows that if the February sample of 194 clusters had been optimally allocated, precision would have improved from $\pm$7.3% to $\pm$6.5%.

3.2 Effect of Further Stratification on the Ratio-to-Cluster-Size Approach

We divided the four strata in half, yielding eight strata, and repeated the analysis. (Because different sampling rates were used in different strata, the dataset would not easily allow for analyzing strata that cross old stratum boundaries.) The resulting sample sizes necessary to achieve various precision levels, using optimal allocation, are shown in Table 4. The within-strata C.O.V.'s shown were estimated from the February dataset. To attain a $\pm$10% precision, an overall sample of 53 half-runs is called for, a decrease of 35% as compared with the 80 clusters required when using four strata. Some strata are very lightly sampled, however. If we insist at least four half-runs per stratum and optimally allocate otherwise, the result, shown in Table 5, is an overall sample size of 58, which is still 28% better than using four strata.

TABLE 2: Stratified Cluster Sampling
Ratio-to-Cluster-Size Approach
Required Sample Size For Desired Precision

November 1986 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Sample Size Prec.= 5%	Sample Size Prec.= 10%	Sample Size Prec.= 15%	Sample Size Prec.= 20%
1	100	7574	929	4.1	52	117.9	0.26	98	24	11	6
2	50	5531	586	4.7	54	74.6	0.34	59	15	7	4
3	0	1584	125	6.3	44	30.0	0.66	13	3	1	1
4	express	2800	508	2.8	48	47.6	0.93	52	13	6	3
						Total Number of Clusters:		222	55	25	14
						Expected Number of Trips:		904	224	102	57
						COV of Total Boardings:		0.024	0.048	0.071	0.095
						95% Precision:		5.0%	10.1%	15.0%	20.0%

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Sample Size Prec.= 5%	Sample Size Prec.= 10%	Sample Size Prec.= 15%	Sample Size Prec.= 20%
1	100	7507	937	4.0	49	111.8	0.32	145	36	16	9
2	50	5535	589	4.7	53	68.0	0.45	91	23	10	6
3	0	1504	126	6.0	44	29.8	0.70	17	4	2	1
4	express	2823	580	2.4	48	48.4	0.94	69	17	8	4
						Total Number of Clusters:		322	80	36	20
						Expected Number of Trips:		1277	317	143	79
						COV of Total Boardings:		0.024	0.048	0.071	0.096
						95% Precision:		5.0%	10.0%	15.0%	20.1%

TABLE 3: Stratified Cluster Sampling
Ratio-to-Cluster-Size Approach
Precision Achieved with Alternative Sample Sizes

November 1986 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	-50% n = 99	-25% n = 149	-10% n = 178	Observed Sample n = 198	+10% n = 218
1	100	7574	929	4.1	52	117.9	0.26	44.0	66.0	78.0	87.0	96.0
2	50	5531	586	4.7	54	74.6	0.34	26.0	40.0	47.0	53.0	58.0
3	0	1584	125	6.3	44	30.0	0.66	6.0	9.0	11.0	12.0	13.0
4	express	2800	508	2.8	48	47.6	0.93	23.0	35.0	42.0	46.0	51.0
						Total Number of Clusters:		99	150	178	198	218
						Expected Number of Trips:		403	611	725	806	888
						COV of Total Boardings:		0.036	0.029	0.027	0.025	0.024
						95% Precision:		7.5%	6.1%	5.6%	5.3%	5.0%

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	-50% n = 97	-25% n = 146	-10% n = 175	Observed Sample n = 194	+10% n = 213
1	100	7507	937	4.0	49	111.8	0.32	44.0	66.0	79.0	87.0	96.0
2	50	5535	589	4.7	53	68.0	0.45	27.0	41.0	50.0	55.0	60.0
3	0	1504	126	6.0	44	29.8	0.70	5.0	8.0	9.0	10.0	11.0
4	express	2823	580	2.4	48	48.4	0.94	21.0	31.0	38.0	42.0	46.0
						Total Number of Clusters:		97	146	176	194	213
						Expected Number of Trips:		385	579	698	770	845
						COV of Total Boardings:		0.043	0.035	0.032	0.031	0.029
						95% Precision:		9.1%	7.4%	6.8%	6.5%	6.2%

TABLE 4: Stratified Cluster Sampling
Ratio-to-Cluster-Size Approach
Effect of Further Stratification

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Observed Prec.= 5.0%	Prec.= 6.8%	Prec.= 10.0%	Prec.= 15.0%	Prec.= 20.0%
1	0	938	77	6.1	23	23.98	0.713	7	4	2	1	0
2	35	1191	114	5.2	22	42.82	0.252	6	3	1	1	0
3	50	2714	249	5.4	23	62.05	0.484	35	19	9	4	2
4	75	4051	470	4.3	33	88.93	0.327	50	27	13	6	3
5	100	3074	377	4.1	19	111.00	0.242	35	19	9	4	2
6	115	2832	345	4.1	18	127.95	0.359	56	30	14	6	3
7	exp 0	662	122	2.7	18	28.47	0.582	5	3	1	1	0
8	exp 50	2005	344	2.9	27	46.85	0.405	16	9	4	2	1
					Total Number of Clusters:			210	114	53	23	13
					Expected Number of Trips:			913	493	228	101	57
					COV of Total Boardings:			0.024	0.032	0.048	0.071	0.095
					95% Precision:			5.0%	6.8%	10.0%	15.0%	20.0%

TABLE 5: Stratified Cluster Sampling
Ratio-to-Cluster-Size Approach
Effect of Further Stratification with Minimum Sample Size (4)

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Optimal @ 10%	Adjusted (min. 4)
1	0	938	77	6.1	23	23.98	0.713	2	4
2	35	1191	114	5.2	22	42.82	0.252	1	4
3	50	2714	249	5.4	23	62.05	0.484	9	8
4	75	4051	470	4.3	33	88.93	0.327	13	12
5	100	3074	377	4.1	19	111.00	0.242	9	9
6	115	2832	345	4.1	18	127.95	0.359	14	13
7	exp 0	662	122	2.7	18	28.47	0.582	1	4
8	exp 50	2005	344	2.9	27	46.85	0.405	4	4
					Total Number of Clusters:			53	58
					Expected Number of Trips:			228	252
					COV of Total Boardings:			0.048	0.047
					95% Precision:			10.0%	10.0%

3.3 Stratified Cluster Sampling With Selection Probability Proportional to Size

Table 6 shows the results of the selection-probability proprotional-to-size approach based on the November and February datasets using the standard four strata on weekdays only. Again, the February data show more variability. The sample sizes shown in Table 6 are those used in November and February, and yield attained precisions of ±6.0% and ±7.3%, which is about the same as the ratio-to-cluster-size approach.

It is also worth noting the difference in the stratum estimates of total daily boardings obtained by the correct expansion procedure (using an unweighted average of cluster trip-level averages) versus the incorrect expansion procedure of dividing the sample total by the fraction of trips sampled. Table 7 displays the results.

Table 8 shows sample size necessary to achieve various precisions, using statistical inputs from the February dataset, and with optimal allocation between strata (to minimize overall number of clusters). Achieving ±10% precision requires a sample of 66 half-runs overall, 18% less than required by the ratio-to-cluster-size approach.

However, the clusters sampled in the selection-probability-proportional to cluster-size approach will tend to be larger clusters than average, and consequently the number of trips called for by this approach is actually 32% greater than the ratio-to-cluster size approach with its 80 required clusters. Again, precision is robust with respect to departures from optimal allocation. If equal sampling rates are used instead of optimal allocation for 66 half-runs sampled, precision worsens from ±10% to ±10.6%.

Further stratifying the sample data, into eight strata as in the previous section, did not yield the significant decrease in required sample size observed in the ratio-to-cluster-size approach. Sample sizes necessary to achieve various precision

TABLE 6: Stratified Cluster Sampling
Selection Probability Proportional to Cluster Size
Precision Achieved with Sampled Data

November 1986 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Expected Sample Cluster Size	Sampled Clusters	Mean Brdngs	Per Cluster Squared COV	Sample Size
1	100	7574	929	4.8	52	118.0	0.084	52
2	50	5531	586	5.7	54	74.3	0.111	54
3	0	1584	125	11.7	44	33.2	0.216	44
4	express	2800	508	4.1	48	43.5	0.397	48

Total Number of Clusters:	198
Expected Number of Trips:	1048
COV of Total Boardings:	0.029
95% Precision:	6.0%

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Expected Sample Cluster Size	Sampled Clusters	Mean Brdngs	Per Cluster Squared COV	Sample Size
1	100	7507	937	4.9	49	118.2	0.121	49
2	50	5535	589	5.9	53	72.8	0.175	53
3	0	1504	126	9.1	44	33.1	0.177	44
4	express	2823	580	3.4	48	42.8	0.288	48

Total Number of Clusters:	194
Expected Number of Trips:	1003
COV of Total Boardings:	0.035
95% Precision:	7.3%

TABLE 7: Stratified Cluster Sampling
Comparison of Systemwide Boardings Using
Ratio and Proportional-to-Cluster-Size (PPS) Methods

November 1986 - Weekday

Stratum	Thrshld	Total Trips	Boardings/Trip Ratio	Boardings/Trip PPS	System Boardings Ratio	System Boardings PPS	Diff	% Diff
1	100	7574	117.9	118.0	893278	893732	-454	-0.1%
2	50	5531	74.6	74.3	412613	410843	1770	0.4%
3	0	1584	30.0	33.2	47441	52573	-5132	-10.8%
4	express	2800	47.6	43.5	133140	121856	11284	8.5%
					1486471	1479004	7467	0.5%

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Boardings/Trip Ratio	Boardings/Trip PPS	System Boardings Ratio	System Boardings PPS	Diff	% Diff
1	100	7507	111.8	118.2	839433	887252	-47820	-5.7%
2	50	5535	68.0	72.8	376214	403114	-26900	-7.2%
3	0	1504	29.8	33.1	44864	49812	-4948	-11.0%
4	express	2823	48.4	42.8	136690	120768	15922	11.6%
					1397201	1460947	-63746	-4.6%

TABLE 8: Stratified Cluster Sampling
Selection Probability Proportional to Cluster Size
Required Sample Size For Desired Precision

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Expected Sample Cluster Size	Sampled Clusters	Mean Brdngs	Per Cluster Squared COV	Sample Size Prec.= 5%	Sample Size Prec.= 10%	Sample Size Prec.= 15%	Sample Size Prec.= 20%
1	100	7507	937	5.8	49	118.2	0.121	144.0	36.0	16.0	9.0
2	50	5535	589	7.1	53	72.8	0.175	78.0	20.0	9.0	5.0
3	0	1504	126	10.3	44	33.1	0.177	10.0	2.0	1.0	1.0
4	express	2823	580	4.0	48	42.8	0.288	30.0	8.0	3.0	2.0
						Total Number of Clusters:		262	66	29	17
						Expected Number of Trips:		1618	408	179	105
						COV of Total Boardings:		0.024	0.047	0.072	0.094
						95% Precision:		5.0%	10.0%	15.0%	19.7%

levels, using optimal allocation, are shown in Table 9. The within-strata C.O.V.'s shown were estimated from the February dataset. To attain a ±10% precision, an overall sample of 60 half-runs is called for, a decrease of only 9% as compared with the 66 clusters required when using four strata. Again, some strata are very lightly sampled, and, if we insist at least four half-runs per stratum and optimally allocate otherwise, the result, shown in Table 10, is an overall sample size of 64, which is only 3% better than using four strata.

3.4 Revenue-Based Estimation

The first analysis done for the revenue-based approach was to estimate the systematic error in using assumed revenue as surrogate for actual revenue. We estimated revenue using the assumed revenue formula from fare check data, expanding it according to the procedure for sample selection probability proportional to cluster size. This was repeated for six fare checks (from November, 1985 to February, 1987). Comparing with actual revenue during the same six weeks, assumed revenue was found to be 6% above actual. (This implies that the average value of a non-pass boarding is over estimated in SCRTD's current procedure, making partronage estimates about 6% lower than they should be). In all results reported hereafter, we have altered the assumed revenue formula by multiplying it by .9436.

The next step was estimating the relative bias of actual revenue as a surrogate for assumed revenue. Because the assumed revenue estimates are based on a sample, they have some sampling variance. Any variation between assumed revenue and actual revenue that exceed the expected variation due to sampling would be attributed to bias. Based on the six fare survey weeks, we found less overall variation than would have been expected from sampling variation alone. Our estimate of the relative bias (as defined in Appendix A) was negative and close enough to zero to support the hypothesis of no

TABLE 9: Stratified Cluster Sampling
Selection Probability Proportional to Cluster Size
Effect of Further Stratification

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Sample Size
1	0	938	77	10.7	23	29.1	0.224	23
2	35	1191	114	7.3	22	41.9	0.072	22
3	50	2714	249	8.0	23	71.3	0.176	23
4	75	4051	470	6.1	33	90.9	0.072	33
5	100	3074	377	6.0	19	108.3	0.147	19
6	115	2832	345	6.1	18	134.7	0.146	18
7	exp 0	662	122	4.7	18	30.3	0.348	18
8	exp 50	2005	344	4.7	27	51.7	0.196	27
					Total Number of Clusters:			183
					Expected Number of Trips:			1165
					COV of Total Boardings:			0.035
					95% Precision:			7.4%

TABLE 10: Stratified Cluster Sampling
Selection Probability Proportional to Cluster Size
Effect of Further Stratification with Minimum Sample Size (4)

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Bus Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Optimal @ 10%	Adjusted (min. 4)
1	0	938	77	10.7	23	29.1	0.224	1.0	4
2	35	1191	114	7.3	22	41.9	0.072	2.0	4
3	50	2714	249	8.0	23	71.3	0.176	9.0	8
4	75	4051	470	6.1	33	90.9	0.072	11.0	10
5	100	3074	377	6.0	19	108.3	0.147	14.0	13
6	115	2832	345	6.1	18	134.7	0.146	17.0	16
7	exp 0	662	122	4.7	18	30.3	0.348	1.0	4
8	exp 50	2005	344	4.7	27	51.7	0.196	5.0	5
					Total Number of Clusters:			60	64
					Expected Number of Trips:			382	408
					COV of Total Boardings:			0.047	0.048
					95% Precision:			9.9%	10.0%

significant bias, meaning that overpayments and underpayments tend to balance in a nearly identical way each month. Our later anlaysis are based therefore on zero bias.

Table 11 summarizes the calculations from which the adjustment factor and bias were estimated.

The November, 1986 and February, 1987 datasets were then analysed with respect to revenue-based estimation. The per-cluster C.O.V. of the conversion factor was found to be .37 in November and .38 in February. Again, the February dataset showed more variation, and was used in our further analyses.

Table 12 shows the number of half-runs to be sampled to attain various precision levels. For a ±10% precision, the sample size needed is 59. This is a reduction of 26% compared with the 80 half-runs called for by the ratio-to-cluster-size approach and 11% compared with the selection-probability-proportional-to-size approach.

3.5 Ex-Post Facto Stratification

For this approach, clusters are stratified at the line/direction/time period (L/D/TP) level using mean boardings per trip on each L/D/TP as derived from the file of most recent our ride checks. Local and express lines were not kept separate in analyses. Time periods were defined as follows:

a.m. peak	6:00am - 9:30am
base	9:00am - 2:00pm
p.m. peak	2:00pm - 6:30pm
evening	6:30pm - midnight
owl	midnight - 6:00am

Each trip was classified according to its start time.

Table 11: Revenue-Based Estimation
Calculation of Bias and Revenue Adjustment Factors

Revenue Adjustment

Month	Actual Revenue	Assumed Revenue	Actual/ Assumed
Aug-85	379881	418748	0.9072
Nov-85	389672	396060	0.9839
Mar-86	383090	427003	0.8972
Jun-86	376841	397421	0.9482
Nov-86	379570	395129	0.9606
Feb-87	380477	394408	0.9647
		mean	0.9436

Bias Calculation

Month	Actual Assumed Revenue	Adjusted Assumed Revenue	Squared Error(SE)	Variance SysRev(S)	SE-var(S)
Aug-85	379881	395140	232837081	398907479	-166070398
Nov-85	389672	373731	254115481	243973621	10141860
Mar-86	383090	402929	393585921	307961676	85624245
Jun-86	376841	375015	3334276	225245649	-221911373
Nov-86	379570	372853	45118089	143632444	-98514355
Feb-87	380477	372172	68973025	197762034	-128789009
Totals	2289531				-519519032

Bias factor (b^2)=-0.0005946

Table 12: Revenue-Based Estimation
Require Sample Sizes Based on February 1987 Data

COV = 0.38

Precision	5%	10%	15%	20%
Sample	233	59	27	16

Our first analysis used four strata of approximately equal size, with boardings thresholds of 30, 50, and 80 separating the strata. The November '86 and February '87 datasets were analyzed. Of particular interest was the size of the inter-cluster correlation effect (called V_2 in Appendix A) relative to the intra-cluster variance (V_1). (The intra-cluster variance is the usual variance arising in stratified cluster sampling; the inter-clusters effect arises from stratifying after sample selection). We expected V_2 to be small compared to V_1, especially as the number of strata increased.

With 4 strata, the November '86 data show V_2 to be -3% of V_1. (The negative sign implies a resulting reduction in sample size requirement because, within a sampled half-run, a L/D/TP cluster that is in the upper half of its stratum tends to be balanced by a L/D/TP cluster in the bottom half of another stratum. However, the small value of V_2 more than likely supports a hypothesis of zero inter-cluster effect.) In February '87, however, V_2 was 12% of V_1, implying a significant inter-cluster effect. To be conservative, we based our further analyses on the February data.

To obtain a $\pm$10% precision, with every half-run in the system selected with equal probability, the L/D/TP stratification approach with four strata requires a sample of 74 half-runs, a 8% improvement over the 80 half-runs required by line stratification with four strata. Table 13 supplies further details. We also tried using 8 strata, but the results were worse.

3.6 Direct Stratification of Clusters

The dataset of recent past ride checks was used to compute an average boardings per trip for each L/D/TP, where time periods for this purpose are 1-hr periods. The current schedule was then analyzed by assigning to each trip in the schedule an expected boardings, which was the average of that trip's L/D/TP as just computed from the ride check database. The current schedule was then split into half-runs, and average expected boardings per trip was calculated for each half-run.

Table 13: Ex Post Facto Stratification
By Line/Direction/Time Period

November 1986 - Weekday

SCHEDULE DATA
Total Super Clusters= 4342

Stratum	Threshold	Trips	Super Clusters	% SC w/ Trip
1	0	1712	761	18%
2	30	2781	1331	31%
3	50	4345	2143	49%
4	80	8170	2702	62%

% Super-Clusters with Trips in Stratum X & Y

Stratum	2	3	4
1	11%	9%	3%
2		15%	11%
3			30%

FARE CHECK DATA

Number of Super Clusters w/ Trips in Stratum X & Y

Stratum	2	3	4
1	20	12	0
2		24	2
3			20

Stratum	COV	Trips/ Cluster	Bdgs/ Trip	V
1	0.79	3.19	35.61	89.74
2	0.59	2.16	63.70	81.18
3	0.41	2.38	94.86	92.56
4	0.39	2.30	112.57	100.98

Limited Correlation Coefficient (Hh)

Stratum	2	3	4
1	0.0805	-0.3845	
2		0.2003	-1.3574
3			0.9878

Relative Covariance	System Boardings Stratum	
-0.0316	1	60964
(V2/V1)	2	177150
	3	412167
	4	919697
	Total	1569978

U_x^2 = 0.1232

Precision	5%	10%	15%	20%
Sample	217	54	24	14

February 1987 - Weekday

SCHEDULE DATA
Total Super Clusters= 4342

Stratum	Threshold	Trips	Super Clusters	% SC w/ Trip
1	0	1712	761	18%
2	30	2781	1331	31%
3	50	4345	2143	49%
4	80	8170	2702	62%

% Super-Clusters with Trips in Stratum X & Y

Stratum	2	3	4
1	11%	9%	3%
2		15%	11%
3			30%

FARE CHECK DATA

Number of Super Clusters w/ Trips in Stratum X & Y

Stratum	2	3	4
1	36	12	0
2		25	2
3			21

Stratum	COV	Trips/ Cluster	Bdgs/ Trip	V
1	0.63	3.21	39.25	79.38
2	0.68	2.35	54.61	87.27
3	0.44	2.38	90.80	95.09
4	0.43	2.38	104.88	107.33

Limited Correlation Coefficient (Hh)

Stratum	2	3	4
1	0.7475	-0.1176	
2		0.7312	0.0422
3			0.7445

Relative Covariance	System Boardings Stratum	
0.1210	1	67196
(V2/V1)	2	151870
	3	394526
	4	856870
		1470462

U_x^2 = 0.1681

Precision	5%	10%	15%	20%
Sample	297	74	33	19

Stratification thresholds of average expected boardings were selected so as to divide the clusters into 8 strata of nearly equal size. Because there was no ride check data available for classifying some clusters, a ninth stratum had to be created for them, containing about 3 percent of the half-runs in the population and about 15 percent of the half-runs in the fare check database. From the schedule database thus modified, we calculated population figures such as the number of trips in each stratum.

The fare check databases were then analyzed according to the ratio-to-cluster size approach. Table 14 shows the number of clusters sampled in each stratum with the stratum C.O.V. (on a per cluster basis), and the resulting precision, for both November 1986 and February 1987. Table 15 shows the sample size needed to obtain various precision levels based on the February 1987 dataset (which requires higher sample sizes than the November 1986 dataset). The number of half-runs to be sampled to meet the 10% precision requirement is 34. In Table 16 the sample sizes have been adjusted to require a minimum of 4 clusters in each stratum, increasing the total sample to 38.

It may be possible to improve the efficiency of this approach by classifying trips which currently cannot be placed in a stratum because no prior ridecheck data is available. This may be done by increasing the length of the time period (1 hour was used), assigning these trips to the closest time period for which ridecheck data are available, or some other method. However, as long as the "unclassified" stratum remains small, little benefit will be derived from reducing its size or eliminating it.

4. Summary/Recommendations

Of the six approaches examined for estimating monthly patronage, the use of direct stratification with the ratio-to-cluster-size approach is recommended. This approach was the most efficient (requiring the smallest sample size to

TABLE 14: Direct Stratification of Clusters
Ratio-to-Cluster-Size Approach
Precision Achieved with Sample Data

November 1986 - Weekday

Stratum	Thrshld	Total Trips	Half Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Sample Size
0	missing	411	134	3.1	30	37.2	0.960	30
1	0	2703	639	4.2	52	29.5	0.725	52
2	40	1966	429	4.6	23	51.0	0.254	23
3	55	2393	544	4.4	16	63.6	0.177	16
4	70	2823	644	4.4	22	90.6	0.219	22
5	85	2381	556	4.3	15	83.8	0.139	15
6	100	1634	391	4.2	12	119.3	0.199	12
7	110	2304	608	3.8	16	120.8	0.240	16
8	130	1370	397	3.5	11	137.2	0.186	11
					Total Number of Clusters:			197
					Expected Number of Trips:			804
					COV of Total Boardings:			0.020
					95% Precision:			4.2%

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Half Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Sample Size
0	missing	411	134	3.1	14	28.3	1.664	14
1	0	2703	639	4.2	52	30.0	0.506	52
2	40	1966	429	4.6	35	44.6	0.352	35
3	55	2393	544	4.4	21	69.4	0.253	21
4	70	2823	644	4.4	25	76.3	0.247	25
5	85	2381	556	4.3	16	117.3	0.190	16
6	100	1634	391	4.2	11	107.1	0.197	11
7	110	2304	608	3.8	9	128.6	0.313	9
8	130	1370	397	3.5	10	142.3	0.281	10
					Total Number of Clusters:			193
					Expected Number of Trips:			780
					COV of Total Boardings:			0.028
					95% Precision:			5.9%

TABLE 15: Direct Stratification of Clusters
Ratio-to-Cluster-Size Approach
Required Sample Size for Desired Precision

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Half Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Prec.= 5.0%	Prec.= 5.9%	Prec.= 10.0%	Prec.= 15.0%	Prec.= 20.0%
0	missing	411	134	3.1	14	28.3	1.664	6	5	2	1	0
1	0	2703	639	4.2	52	30.0	0.506	13	10	3	1	1
2	40	1966	429	4.6	35	44.6	0.352	10	7	3	1	1
3	55	2393	544	4.4	21	69.4	0.253	14	10	3	2	1
4	70	2823	644	4.4	25	76.3	0.247	17	13	4	2	1
5	85	2381	556	4.3	16	117.3	0.190	17	12	4	2	1
6	100	1634	391	4.2	11	107.1	0.197	11	8	3	1	1
7	110	2304	608	3.8	9	128.6	0.313	30	22	8	3	2
8	130	1370	397	3.5	10	142.3	0.281	18	13	4	2	1
					Total Number of Clusters:			138	99	34	15	9
					Expected Number of Trips:			558	401	139	62	35
					COV of Total Boardings:			0.024	0.028	0.048	0.071	0.095
					95% Precision:			5.0%	5.9%	10.0%	15.0%	20.0%

TABLE 16: Direct Stratification of Clusters
Ratio-to-Cluster-Size Approach
Effect of Minimum Sample Size (4)

February 1987 - Weekday

Stratum	Thrshld	Total Trips	Half Runs	Mean Cluster Size	Sampled Clusters	Mean Brdngs	COV	Optimal @ 10.0%	Adjusted (min. 4)
0	missing	411	134	3.1	14	28.3	1.664	2	4
1	0	2703	639	4.2	52	30.0	0.506	3	4
2	40	1966	429	4.6	35	44.6	0.352	3	4
3	55	2393	544	4.4	21	69.4	0.253	3	4
4	70	2823	644	4.4	25	76.3	0.247	4	4
5	85	2381	556	4.3	16	117.3	0.190	4	4
6	100	1634	391	4.2	11	107.1	0.197	3	4
7	110	2304	608	3.8	9	128.6	0.313	8	6
8	130	1370	397	3.5	10	142.3	0.281	4	4
					Total Number of Clusters:			34	38
					Expected Number of Trips:			139	154
					COV of Total Boardings:			0.048	0.047
					95% Precision:			10.0%	9.9%

attain a given precision) and would be relatively straightforward to implement.

Based on February 1987 data, an estimated 38 clusters would be required to achieve 10% precision for week day patronage.

It should be noted, however, that this low sample size is due in part to the accuracy of recent ride check data which are the basis for stratification. If ride checks are done less frequently, or are estimated using point checks with some loss of accuracy, accuracy of the patronage estimate may suffer. To provide a margin of safety during the first year of implementation, then, we recommend that 52 half-runs be sampled per quarter, sampling one half-run on four days a week for 13 weeks. The sampling days each week should be selected at random. This level of effort is about 75 percent less than the level of fare check sampling now being done.

We recommend sampling weekends at approximately the same level of intensity, i.e., one half-run each day. The sample size for a given day (Saturday or Sunday) will be four times smaller than the weekday sample size, meaning quarterly estimates will be quite unreliable; however, annual estimates should have a 10% precision.

Some new software will be needed to implement the new sampling plan. Most of it can be extracted from the programs we have developed to test the plan. Programs are needed to (1) process the file of recent ride checks to determine average boardings by L/D/TP, and attach that estimate to every trip in the schedule file; (2) process the schedule file by forming clusters, estimating the expected mean boardings per trip in each cluster, and assigning it to a stratum; (3) select clusters at random from within each stratum; and (4) process the fare check data to calculated mean boardings per trip in each stratum, and expand it to a population total. SCRTD may also want to recalculate the per-cluster C.O.V.'s every year

using the data of the previous year (one quarter's worth of data is not enough), and revise sample sizes if necessary. The software for calculating sample sizes from per-cluster C.O.V.'s may be copied from our LOTUS program.

Regardless of which approach is used by RTD, those changes outlined in Section 1.2 that have not yet been adopted should be implemented. These include: including trippers and interlined runs in the sampling frame; presplitting runs and selecting pieces (i.e., half-runs); spreading out the data collection over the entire year; and taking missed trips into account when expanding the sample.

0351P

APPENDIX A

SAMPLE SIZE AND PRECISION FORMULAS

1. Notation

h = stratum index

i = cluster index

j = trip index

M_h = number of trips in cluster i

y_{hij} = boardings on trip j of cluster i

$y_{hio} = \sum_{j=1}^{M_{hi}} y_{hij}$ = total boardings in cluster i

$\bar{y}_{hi} = y_{hio}/m_{hi}$ = trip mean boardings for cluster i

N_h = number of stratum h cluster in population

n_h = number of stratum h clusters in sample

M_{ho} = number of stratum h trips in population

m_{ho} = number of stratum h trips in sample

$\bar{M}_h = M_{ho}/N_h$ = mean cluster size in stratum h

p_h = number of stratum h clusters in analysis dataset

z = z-value corresponding to confidence level (e.g., $z = 1.96$ for 95% confidence level when standard deviation is known)

d = precision (e.g., $d = 0.1$ means $\pm 10\%$)

2. Ratio-to-Cluster-Size Estimators for Stratified Cluster Sampling

This derivation applies to both the approach of stratifying lines and the approach of directly stratifying clusters. Using the ratio-to-cluster-size approach (Cochran, section 9.8), the stratum h boardings per trip estimator is

$$\bar{y}_h = \left[\sum_{i=1}^{n_h} \sum_{j=1}^{M_{hi}} y_{hij} \right] / m_{ho} \qquad (A.1)$$

The estimator of stratum total boardings is $Y_h = M_{ho}\bar{y}_h$, and the system total estimator is

$$Y_o = \sum^{h} \bar{y}_h M_{ho} = \sum^{h} Y_h \qquad (A.2)$$

These estimators are unbiased if every cluster in a given stratum has an equal chance of being selected.

2.1 Stratum Variance

The variance of the stratum total estimator is estimated to be

$$V(Y_h) = \frac{N_h^2}{n_h} \frac{\sum_{i=1}^{p_h} (y_{hio} - M_{hi} \bar{y}_h)^2}{p_h - 1} \qquad (A.3)$$

The squared C.O.V. (coefficient of variation) of Y_h, which is also the squared C.O.V. of the mean boardings per trip, is

$$v^2(Y_h) = V(Y_h) \; / \; (N_h \bar{M}_h \bar{y}_h)^2 \qquad (A.4)$$

The squared C.O.V. on a per cluster basis is defined to be

$$u_h^2 = n_h \, v^2(Y_h) = \frac{1}{\bar{M}_h^2 \bar{y}_h^2} \frac{\sum_{i=1}^{p_h} (y_{hio} - M_{hi} \bar{y}_h)^2}{p_h - 1} \qquad (A.5)$$

The per cluster C.O.V. term is convenient since it can be calculated without prior specification of n_h, and because it can be applied in the well-known formula

$$\text{C.O.V. of mean or total} = \frac{\text{C.O.V. of sampling unit}}{\sqrt{\text{number of units sampled}}}$$

2.2 System Total Variance

The system total variance is simply

$$V(Y_o) = \sum_h V(Y_h) \qquad (A.6)$$

since the clusters in each stratum are selected independently. The system total C.O.V. is therefore

$$v(Y_o) = \frac{\sqrt{\sum u_h^2 M_{ho}^2 \bar{y}_h^2 / n_h}}{\sum M_{ho} \bar{y}_h} \qquad (A.7)$$

and its precision (relative tolerance) is

$$d = z \, v(Y_o) \qquad (A.8)$$

2.3 Optimal Allocation Between Strata

To minimize the C.O.V. for a given overall sample size, or to minimize overall sample size for a desired overall C.O.V., the number of clusters

sampled in stratum h should be proportional to $u_h M_{ho} \bar{y}_h$, under the assumption that sampling cost is proportional to number of clusters sampled. This is easily seen by minimizing $v(Y_o)$ subject to a constraint on Σn_h, or by minimizing Σn_h subject to a given $v(Y_o)$.

For a given number of clusters n_o, optimal allocation is

$$n_h = n_o \frac{u_h M_{ho} \bar{y}_h}{\Sigma u_h M_{ho} \bar{y}_h} \tag{A.9}$$

and for a given desired precision d,

$$n_h = u_h M_{ho} \bar{y}_h \frac{\Sigma u_h M_{ho} \bar{y}_h}{\left(\frac{d}{z}\right)^2 \left(\Sigma M_{ho} \bar{y}_h\right)^2} \tag{A.10}$$

However, if cost is proportional to the number of trips sampled (rather than number of runs), it follows that strata with more trips per run should be sampled less. Optimizing leads to

$$n_h \propto u_h \bar{y}_h N_h \sqrt{\bar{M}_h} \tag{A.11}$$

Most generally, if the cost of sampling a stratum h cluster is c_h, optimum allocation calls for

$$n_h \propto u_h \bar{y}_h M_{ho} / \sqrt{c_h} \tag{A.12}$$

3. Stratified Cluster Sampling with Probability Proportional to Size

Another sampling approach that recognizes differing cluster sizes is to sample clusters with probability proportional to the number of trips in the cluster (Cochran, sections 9.9 - 9.10). The unbiased estimator of stratum h boardings is

$$Y_{hpps} = \frac{M_{ho}}{n_h} \sum_{i=1}^{n_h} \bar{y}_{hi} = M_{ho} \bar{\bar{y}}_h \tag{A.13}$$

where $\bar{\bar{y}}_h$ is the unweighted mean of the cluster trip-level means, and does not equal $\bar{y}_h$ as defined in equation (A.1).

The system total estimator is $Y_{opps} = \sum^{h} Y_{hpps}$

3.1 Stratum Variance

The variance of the stratum total estimator is estimated to be

$$V(Y_{hpps}) = \frac{M_{ho}^2}{n_h (p_h - 1)} \sum_{i=1}^{p_h} (\bar{y}_{hi} - \bar{y}_h)^2 \qquad (A.14)$$

Its squared C.O.V. on a per cluster basis is

$$u_{hpps}^2 = n_h v_{hpps}^2 = \frac{\sum_{i=1}^{p_h} (\bar{y}_{hi} - \bar{y}_h)^2}{(p_h - 1) \bar{y}_h^2} \qquad (A.15)$$

where v_{hpps} is the C.O.V. of the stratum total.

3.2 System Total Variance

The system total variance is again the sum of the strata total variances;

$$V(Y_{opps}) = \sum_{h} V(Y_{hpps}) \qquad (A.16)$$

The system total C.O.V. is

$$v(Y_{opps}) = \frac{\sqrt{\sum_{h} u_{hpps}^2 M_{ho}^2 \bar{y}_h^2 / n_h}}{\sum_{h} M_{ho} \bar{y}_h} \qquad (A.17)$$

Its precision is $d = z\, v(Y_{opps})$.

3.3 Optimal Allocation Between Strata

If the sampling cost is proportional to number of runs sampled, and all strata have the same cost per run, optimal allocation between strata calls for

$$n_h \propto u_{hpps} M_{ho} \bar{y}_h \qquad (A.18)$$

If cost is proportional to number of trips, it is important to recognize that the expected number of trips per cluster sampled in stratum h is greater than M_h since bigger clusters are sampled with greater probability. The expected size of a sampled cluster is $\bar{M}_h(1 + v_{mh}^2)$, where v_{mh} is the C.O.V. of cluster size in stratum h. Therefore optimum allocation is

$$n_h \propto u_{hpps} \bar{y}_h N_h \sqrt{\bar{M}_h / (1 + v_{Mh}^2)} \qquad (A.19)$$

4. Unstratified Ratio-to-Revenue Cluster Sampling

4.1 Further Notation

s_{hij} = assumed (calculated) revenue on trip j of cluster i

$s_{hi} = s_{hij}/M_{hi}$ = trip mean assumed revenue in cluster i

4.2 Elimination of Systematic Bias From Assumed Revenue Formula and Calculation of Random Bias

Since the analysis datasets have clusters selected with probability proportional to size, the unbiased estimator of daily systemwide assumed revenue during a fare check week is

$$S_{po} = \sum \frac{M_{ho}}{p_h} \sum_{i=1}^{p_h} \bar{s}_{hi} = \sum^{h} M_{ho} \bar{s}_h \qquad (A.20)$$

and its variance is estimated as

$$V(S_{po}) = \sum^{h} \frac{M_{ho}^2}{p_h (p_h - 1)} \sum^{p_h} (\bar{s}_{ih} - \bar{s}_h)^2 \qquad (A.21)$$

By calculating S_{po} for several fare check weeks and comparing it with actual revenue, we then adjusted the formula for assumed revenue to make it "neutral biased" as an estimator of actual revenue.

With s_{hij} thus adjusted, the relative squared bias in using actual revenue as an estimator of assumed revenue can be estimated by comparing, over several fare survey months, the average squared error to the expected sampling squared error. The relative squared bias is estimated to be

$$\beta^2 = \left[\frac{\sum_{w=1}^{n_w} (S_{ow} - A_{ow})^2}{n_w} - V(S_{ow}) \right] / \bar{S}_o^2 \qquad (A.22)$$

where S_{ow} = mean assumed daily revenue in fare survey week w

A_{ow} = mean actual daily revenue in fare survey week w

$\bar{S}_o$ = mean daily revenue, averaged over all fare survey weeks

n_w = number of fare survey weeks in analysis dataset

The relative bias is taken to be near enough to zero to be negligible if $\beta^2 < 0$, and equal to $(\sqrt{\beta^2})$ otherwise.

4.3 Estimator of Boardings Per Assumed Cash Revenue

(Note: the stratum subscript h is dropped in the remainder of section 4.)

The ratio estimator of the number of boardings per assumed cash revenue is

$$R = \frac{\sum_{i=1}^{n} y_{io}}{\sum_{i=1}^{n} s_{io}} \qquad (A.23)$$

where s_{io} and y_{io} are the cluster i total assumed revenue and boardings.

Systemwide daily boardings is then estimated as

$$Y_R = R\, S_o \qquad (A.24)$$

Note that the conversion factor R is estimated using assumed revenue s_{io}, while it is expanded using actual revenue A_o.

4.4 Variance of the Ratio

The squared C.O.V. of R, v_R^2, and the squared C.O.V. on a per cluster basis, u_R^2, are estimated from an analysis dataset with p clusters as follows:

$$v_R^2 = \frac{u_R^2}{n} = \frac{\sum_{i=1}^{p} (y_{io} - Rs_{io})^2}{n\,(p-1)\,\bar{y}_o^2} \qquad (A.25)$$

where $\bar{y}_o$ is the mean cluster boardings.

Alternately (and equivalently),

$$u_R^2 = \frac{1}{n}\,(v_{ycl}^2 + v_{scl}^2 - 2r_{yscl}\, v_{ycl}\, v_{scl}) \qquad (A.26)$$

where the subscript (cl) denotes a reference to cluster totals, and where $r_{ys(cl)}$ is the correlation coefficient between cluster total boardings and assumed revenue.

4.5 Mean Squared Error of the Total Boardings Estimator

In expanding R, we cannot use S_o (systemwide assumed revenue) because it is unknown; instead we use A_o (systemwide actual revenue). The relative MSE of the systemwide boardings is therefore

$$v^2(Y_R) = \frac{u_R^2}{n} + \beta^2 \qquad (A.27)$$

4.6 Sample Size and Precision

If n clusters are sampled (selected at random, with equal probability), the precision of the system estimate is

$$d = z\, v(Y_R) = z \sqrt{\frac{u_R^2}{n} + \beta^2} \qquad (A.28)$$

The necessary number of clusters to attain a precision level d is

$$n = \frac{u_R^2}{\left(\frac{d}{2}\right)^2 - \beta^2} \qquad (A.29)$$

More precise estimates are unattainable since sample size cannot reduce the error arising from using actual revenue as proxy for assumed revenue.

5. Ex-Post-Facto L/D/TP - Stratified Cluster Sampling

Sampling is done by half-run, as before; now, however, a half-run is called a supercluster. Superclusters are selected without stratification, with equal selection probabilities.

Each trip in the system is labeled with the average boardings per trip of the line/direction/time period (L/D/TP) to which it belongs. These L/D/TP averages come from recent ride checks. Stratification is then done by values of these labels. A cluster is now defined to be the group of trips in a supercluster that lie in the same stratum. A cluster may therefore contain trips of more than one line, direction, or period, if the trips belong to the same supercluster and their labels fall in the same stratum.

5.1 Notation

The indices h and H refer to strata. The index i refers to a cluster. The index k refers to a supercluster.

5.2 Estimators

The estimate of system boardings will be called Y_{ex}. As in Section 2,

$$Y_{ex} = \sum^{h} \bar{y}_h \, M_{ho} \qquad (A.30)$$

The stratum mean boardings per trip estimates are

$$\bar{y}_h = \frac{1}{m_{ho}} \sum_{i=1}^{n_h} \sum_{j=1}^{M_{hi}} y_{hij} \qquad (A.31)$$

5.3 Variance of the Estimate Y_{ex}

Between superclusters, sample selection is random; however, within superclusters, cluster selection is not random, and therefore the variance of the system total must include covariance terms for clusters lying in the same supercluster.

The following derivation omits terms that are $O(n^2)$ and higher.

Since $Y_{ex} = \sum^{h} M_{ho}\, y_h$,

$$E[V(Y_{ex})] = E[\sum^{h} M_{ho}^2\, V(\bar{y}_h)] + 2E[\sum^{h} \sum^{H>h} M_{ho}\, M_{Ho}\, \mathrm{Cov}\,(\bar{y}_h, \bar{y}_H)]$$

$$= E[V_1] + E[V_2] \qquad (A.32)$$

Here, V_1 is the intracluster variance and V_2 is the inter-cluster (but intra-supercluster) contribution to variance.

The intracluster variance V_1 is calculated as in the ratio-to-cluster-size approach (equations (A.3)-(A.5)). Expressing the variance of the stratum h mean in terms of the per cluster C.O.V. yields

$$V(\bar{y}_h) = \frac{u_h^2\, \bar{y}_h^2}{n_h} \qquad (A.32a)$$

Before the sample is selected, n_h is unknown. Using the first order approximation $E(1/n_h) = 1/E(n_h)$, and letting

g_h = fraction of superclusters in the population containing a stratum h trip

we obtain

$$E[V_1] = \sum^{h} M_{ho}^2\, \frac{u_h^2\, \bar{y}_h^2}{n\, g_h} \qquad (A.33)$$

where n = number of superclusters sampled.

References

Cochran, W.G., Sampling Techniques 2nd ed., Wiley, 1963.

0342P

With respect to V_2, the between cluster contribution to variance,

$$\mathrm{Cov}(\bar{y}_h , \bar{y}_H) = \frac{1}{m_{ho}\, m_{Ho}} \mathrm{Cov}\left[\sum_{i=1}^{n_h} y_{hio} , \sum_{i=1}^{n_h} y_{Hio}\right] \qquad (A.34)$$

Then since $y_h = y_{hio}/m_{ho}$, and using the identity $\mathrm{Cov}(\sum_i^i A_i , \sum_j^j B_j) - \sum_i^i \sum_j^j \mathrm{Cov}(A_i , B_j)$,

$$\mathrm{Cov}(\bar{y}_h , \bar{y}_H) = \frac{1}{m_{ho}\, m_{Ho}} \sum_{i=1}^{n_h} \sum_{i'=1}^{n_H} \mathrm{Cov}(y_{hio} , y_{Hi'o}) = n_{hH}\, s_{hH} \qquad (A.35)$$

where s_{hH} is the covariance of a stratum h cluster total with a stratum H cluster total, and n_{hH} is the number of superclusters in the sample that span both strata h and H. The second equality follows because

$$\mathrm{Cov}(y_{hio} , y_{Hi'o}) = \begin{cases} s_{hH} & \text{if cluster i and cluster i' lie in the same supercluster} \\ 0 & \text{otherwise} \end{cases} \qquad (A.36)$$

since superclusters are selected independently. Using the index k to refer to a supercluster, and dropping the indices i and i' (since a supercluster cannot contain more than one cluster in a given stratum), the covariance term is estimated as

$$s_{hH} = \frac{\sum_{k=1}^{p_{hH}} (y_{kho} - M_{kh}\, \bar{y}_h)\,(y_{kHo} - M_{kH}\, \bar{y}_H)}{(p_{hH} - 1)} \qquad (A.37)$$

where

p_{hH} = number of superclusters in the analysis dataset with clusters in both strata h and H.

The corresponding correlation coefficient is

$$r_{hH} = \frac{s_{hH}}{u_h \bar{y}_h \bar{M}_h u_H \bar{y}_H \bar{M}_H} \tag{A.38}$$

We now have

$$V_2 = 2 \sum_{h}^{} \sum_{H>h}^{} \frac{n_{hH}}{m_{ho} m_{Ho}} M_{ho} M_{Ho} r_{hH} u_h \bar{y}_h \bar{y}_H \bar{M}_h \bar{M}_H \tag{A.39}$$

To get $E[V_2]$, we make the first order approximation

$$E\left[\frac{n_{hH}}{m_{ho} m_{Ho}}\right] \doteq \frac{E[n_{hH}]}{E[m_{ho}] E[m_{Ho}]} = \frac{n f_{hH}}{n^2 g_h g_H \bar{M}_h \bar{M}_H} \tag{A.40}$$

where

f_{hH} = fraction of supercluster in the population containing both a stratum h trip and a stratum H trip

Combining the within-cluster and between-cluster effects,

$$E[V(Y_{ex})] = \frac{1}{n}\left\{ \sum_{h}^{} \frac{M_{ho}^2 u_h^2 \bar{y}_h^2}{g_h} + 2 \sum_{h}^{} \sum_{H>h}^{} \frac{M_{ho} M_{Ho} u_h u_H \bar{y}_h \bar{y}_H r_{hH} f_{hH}}{g_h g_H} \right\} \tag{A.41}$$

The per-supercluster squared C.O.V. u_{ex}^2 (which is stratum independent) is given by

$$u_{ex}^2 = \frac{n E[V(Y_{ex})]}{Y_{ex}^2} \tag{A.42}$$

5.4 Sample Size and Tolerance

The tolerance obtained from a sample of n superclusters is

$$d = \frac{z u_{ex}}{\sqrt{n}} \tag{A.43}$$

and the number of superclusters that must be sampled to obtain a tolerance d is

$$n = \left(\frac{z u_{ex}}{d}\right)^2 \tag{A.44}$$