9-1989

Sampling Bus Ridership at the Route Level: Initial Results from Efforts to Improve the Precision of Sample Estimates

James G. Strathman
Portland State University

Sampling Bus Ridership at the Route Level: Initial Results From Efforts to Improve the Precision of Sample Estimates

James G. Strathman

Final Report
September 1989

Center for Urban Studies
School of Urban and Public Affairs
Portland State University
Portland, OR 97207-0751

Dr. Strathman is Assistant Director of the Center for Urban Studies and an Associate Professor of Urban Studies and Planning, Portland State University.

Support for this project was provided by the U.S. Department of Transportation and the Tri-County Metropolitan Transportation District of Oregon (TRI-MET).

Abstract

UMTA's Section 15 reporting requirements establish precision standards and corresponding sampling plans for estimating bus ridership at the system level. However, many transit providers are interested in recovering data with sufficient precision to permit ridership analysis at the route level. One outcome of extending a system sampling plan to route level data collection and analysis would be a large increase in the sample sizes required to achieve a reasonable standard of precision. Given the costs involved in expanding the data recovery process beyond what is required for Section 15 reporting, alternative means of improving the precision of route level data warrant consideration.

This report presents results from an initial effort to estimate ridership per bus trip on the basis of route level characteristics (primarily related to time of service). The analysis covers 17 routes served by the Tri-County Metropolitan Transportation District of Oregon (Tri-Met). We find that variations in ridership across bus trips within each of the routes studied are significantly related to the route characteristics. The effects of these characteristics on ridership were also found to vary from route to route. Improvements in the precision of ridership estimates attributable to the contribution of route-specific information translate, in the limit, to a 45 per cent reduction in the sample size required to achieve a given level of precision. Thus the potential gains in sampling precision from further research on this subject appear promising.

Acknowledgements

The author gratefully acknowledges the efforts of Janet Hopper, who provided considerable input to the formulation of the approach taken in this project. Chall Kyoung Sung also provided vital research and statistical assistance.

Introduction

Public transit providers are required to collect ridership data under UMTA's Section 15 reporting system. The data collected for this purpose, however, concern transit operating characteristics at the system level, whereas many transit planners are more concerned with the evaluation of ridership patterns at the route level. A shift in orientation from the system to the route level has important implications for the design of ridership sampling plans. The composite sample size required to achieve an acceptable level of statistical precision at the route level is substantially larger, and the sampling plan must also be restructured to conform with the relative variations of ridership across routes.

Some transit providers have installed automatic passenger counters (APC's) in their fleets to facilitate collection of large quantities of ridership data, in support of route scheduling and planning activities. The use of APC's introduces another dimension to the sampling exercise, which also affects the data recovery process at the route level [2, 20].

In this report we address several issues associated with route level sampling. First, we discuss the features that distinguish a route level sampling plan from one executed at the system level. Second, we note differences arising from the use of APC's in recovering ridership data. Third, we review alternative approaches to improving the precision of ridership estimates. Lastly, we report the results of an initial application of a clustering methodology designed to improve the precision of ridership estimates at the route level. The route clustering approach uses a classification scheme based on ridership differentials, and is evaluated with respect to reductions in the intra-cluster variation of parameter estimates of the "time of service" determinants of ridership.

Ridership data and other transit operating information used in the analysis were provided by the Tri-County Metropolitan Transit District of Oregon (TRI-MET), which introduced APC's in 1982 and has relied on this technology for Section 15 reporting since 1986.

General Aspects of Route Level Sampling

Procedures for sampling bus trips associated with Section 15 reporting have been defined by UMTA [21]. A principal underlying objective of the sampling plan is to ensure an equal probability of selection among all scheduled bus trips. Ridership estimates for Section 15 reporting are subject to a standard of precision of +/- 10 per cent at the 95 per cent level of confidence. Given this standard and the system level ridership variation, a minimum sample size can be derived as follows [14]:

$$n = [(z_{\alpha/2} \cdot S) / (p \cdot M)]^2 \text{ , where}$$

n = the number of bus trips to be randomly sampled;

$z_{\alpha/2}$ = the critical z value associated with a level of confidence of $(1 - \alpha)$;

S = the standard deviation of ridership per bus trip for the system;

p = the level of precision required from the sample estimate;

M = the mean ridership per bus trip for the system.

Application of the system level sampling plan described above to the route level requires that route specific estimates of mean ridership and its standard deviation be substituted for the system level statistics. The degree of precision sought from route level statistics may also be less rigorous. For example, a standard of +/- 20 per cent at the 90 per cent level of confidence may provide acceptable precision for route level analysis of ridership [17].

Regarding Section 15 reporting, it is important to remember that system level inferences cannot be directly obtained from simple aggregation of the data recovered from a route level sampling plan. In the more disaggregate route sampling approach the probability of selection remains equal for all bus trips within a given route. But because the coefficient of variation on ridership tends to vary across routes, so too does the sampling fraction. Thus post stratification of the sample data by route would be required for inferring system level ridership from a route level sampling plan [15, 20].

The sampling plan, whether oriented to the system or route level, is also affected by the method of data collection. With manual data collection, surveyors can be assigned to individual randomly selected bus trips. Alternatively, APC-equipped vehicles recover ridership data from all the trips served in a daily bus assignment. These blocks of trips - comprising what are termed "trains" - are not independent, as is required for simple random sampling. Rather, the trip level data recovered by APC's constitute a cluster sample where trains represent the primary sampling unit [2, 20]. A sampling plan for APC's must account for the magnifying effects of within-train correlation of bus trips on overall sampling error. The presence of this type of correlation with APC's translates into larger sample size requirements to achieve the same level of precision as a simple random sample, with the proportionate increase in sample size determined by both the level of intra-train correlation and the number of bus trips per train.

The proportionate increase in sample size resulting from the clustering of bus trips within train level selection units has been called the "design effect" [15]. The magnitude of the design effect for a cluster sample of bus trips is defined as follows:

$$Def = [\ 1 + (N - 1) \cdot \rho\ ] \text{ , where}$$

Def = the design effect;

N = the number of bus trips per train;

ρ = the correlation of ridership among bus trips within trains.

We see from this equation that in the extreme instance where there is no intra-train correlation of ridership, the design effect is equal to one, and the cluster and simple random sample sizes are equivalent. Also, we note from the equation that there can be no clustering (or correlation) when N equals one. When either the number of bus trips per train or the intra-train correlation increases, so does the magnitude of the design effect. This is illustrated in Table 1, which presents values of the design effect for alternative train sizes and correlation values. For large trains with relatively strong intra-train correlation, the value of the design effect can be

substantial. For example, given a transit system with APC trains averaging 25 bus trips correlated at .4, the corresponding sample size would need to be nearly eleven times larger than a simple random sample of bus trips.

Table 1

Alternative Values of the Design Effect for a Cluster Sample of Trains

	Intra-train Correlation			
Bus Trips/Train	.2	.4	.6	.8
5	1.8	2.6	3.4	4.2
10	2.8	4.6	6.4	8.2
15	3.8	6.6	9.4	12.2
20	4.8	8.6	12.4	16.2
25	5.8	10.6	15.4	20.2
30	6.8	12.6	18.4	24.2

In designing an APC sampling plan for implementation by Tri-Met, we derived a design effect of 4.3, corresponding with an average train size of 8.98 bus trips and intra-train correlation of .414 [20]. This finding points to one of the basic trade-offs associated with the use of APC's. While they are technically capable of recovering the large number of observations required for route level analysis at a relatively favorable cost per observation [16], they also require larger sample sizes to achieve the same level of precision as a simple random sample with manual data recovery. Cost-benefit studies of alternative data collection methods should thus account for the design effect to avoid a differential "precision bias" favoring APC's over manual data collection.

Conventional Estimates of Route Sample Sizes

To illustrate the effects of a shift in sampling orientation from the system to the route level on the precision of ridership estimates, we evaluated sample data collected by Tri-Met between April and June 1988. We randomly selected 16 routes providing weekday service, 9 routes providing Saturday service and 7 routes providing Sunday service. The routes chosen represent nearly 20 per cent of Tri-Met's service network. Statistics on the number of boarding riders per bus trip for each of the routes were used to determine the precision of the sample estimates at the 90 per cent level of confidence. We then derived the sample size for each route that would be required to achieve precision of +/- 20 per cent at the 90 per cent level of confidence. These results are presented in Table 2.

Table 2

Precision Estimates and Minimum Sample Sizes for Selected Routes: April-June, 1988

Weekday Rts.	Precision (%)[1]	Sample Size[2]
12	37.5	121
57	54.1	128
20	30.1	115
36	125.2	369
89	129.7	177
55	89.9	289
41	28.3	76
38	110.8	209
43	60.9	114
23	65.7	132
120	35.1	155
32	73.7	183
33	60.9	153
60	120.3	82
34	89.7	261
54	41.9	100
Saturday Rts.		
8	86.9	97
108	75.3	184
72	37.0	143
35	46.4	42
109	42.0	89
33	50.6	50
70	64.9	144
141	46.0	77
9	37.5	68
Sunday Rts.		
75	44.6	126
120	61.2	195
70	100.9	124
71	61.9	102
109	58.8	115
4	57.8	49
6	36.7	65

[1] The precision estimates represent the percentage range around the route level sample mean boardings that, with 90 per cent confidence, contains the true population mean.

[2] The sample size is the minimum number of bus trips in each given route that would have to be sampled to achieve +/- 20 per cent precision at the 90 per cent level of confidence. These estimates assume cluster sampling with APC's and a design effect of 4.3.

The level of precision in the sample of routes varies considerably, ranging from 30 per cent on Route 20 (weekdays) to 130 per cent on Route 89 (weekdays). On average, the precision for weekday routes (+/- 72.1 per cent) is less than the precision for Sunday (+/- 60.2 per cent) and Saturday (+/- 54.1 per cent) routes. This is probably the result of the effects of the AM and PM peak ridership "spikes" on weekday variances.

The precision estimates derived at the route level are weak in comparison with the precision of the system level ridership estimate. Previously, we determined that the precision of the system level estimate of ridership during the September-November 1988 period was approximately four per cent at the 95 per cent level of confidence [20].

In order to achieve a reasonable standard of precision, the sample sizes associated with a route level orientation would have to be considerably larger than the sample size required for UMTA Section 15 reporting. Extending the mean sample size estimates from the weekday, Saturday and Sunday routes in Table 2 to all routes in Tri-Met's service network, a sample of 22,815 bus trips would have been required during the April-June 1988 period to achieve 20 per cent precision at the 90 per cent level of confidence. This sample would represent nearly seven per cent of all scheduled trips. By comparison, we determined that a system level sample of .6 per cent of all scheduled bus trips (1,961 trip observations) would satisfy the stricter precision standards pertaining to UMTA Section 15 reporting [20]. Thus the change in orientation from the system to the route level in generating ridership estimates entails more than a ten-fold increase in data recovery requirements, even after allowing for a weaker standard of precision.

Mitigating the concerns regarding such a large increase in sampling requirements at the route level is the acknowledgement that service planning and scheduling cannot function effectively on the basis of system level information. Responsive and cost effective service provision depend on access to transit utilization information at the route level. In light of the

effort and resources required to recover adequate route level data, however, alternative means of improving the precision of disaggregate ridership estimates warrant consideration. Various methods serving this purpose are reviewed below.

Methods for Improving Sample Estimates at the Route Level

At the outset it is important to distinguish between the objective of improving the precision of route level sample estimates and the objective of estimating transit demand at the route level. In essence, with the latter objective it is assumed that the ridership data employed in estimating transit demand is measured without error, whereas with the former objective it is recognized that ridership data is subject to sampling and measurement error [17]. These two objectives are related in the sense that improvements in the precision of route level ridership data resulting from reductions in sampling or measurement error translate into improvements in the performance of models estimating route level transit demand. Elsewhere, we have addressed the issue of measurement error in recovering ridership data with APC's [20]. We will thus limit our attention here to the issue of sampling error.

A wealth of literature exists on the determinants of transit demand at the system level, and attention to estimating demand at the route level has been growing [1, 7, 8, 9, 11, 12, 19]. Typical route level demand models utilize time series data to estimate ridership on the basis of service area characteristics (largely socioeconomic factors associated with trip generation), the level of transit service (represented by platform hours, miles of service or average headways), the relative cost of travel by transit and automobile, traffic congestion and seasonal factors.

Much less attention has been devoted to the objective of reducing sampling error associated with the recovery of transit ridership data. Alternatively, efforts have been made in the area of highway traffic data collection to identify homogeneous subsets of the overall transportation network, wherein data collected for highway links within the subset can be used as a basis for inferring traffic on companion links for which data was not recovered [3]. Given that many state highway networks are comprised of thousands of links (which can be viewed as the

counterparts of transit routes), potentially sizable savings in traffic counting costs could be realized by such a segmentation approach.

One of the means of segmenting highway links has been on the basis of their functional classification, as defined in the Highway Performance Monitoring System [4]. Some improvement in the precision of sample estimates of average annual daily traffic (AADT) has resulted from this type of clustering. Within-cluster reductions of the coefficient of variation for AADT have not been substantial, however, indicating that corresponding reductions in the required sample sizes would be relatively small.

Another alternative that was explored by Gur and Hocherman [6] was to estimate traffic volumes for road links on the basis of their location and physical characteristics, their recorded volumes from previous periods and the current volumes from other links in the cluster. They found that a uniform growth factor applied to previous counts provided better results than cluster-specific growth factors. They also estimated changes in traffic volume for road links on the basis of changes in volume on connected links. This approach may hold less merit in applications to transit ridership, given the relatively weaker ridership inter-relationships between routes in the transit service network [10]. Because transit riders are less inclined to undertake trips involving transfers between routes, a change in ridership on a given route is less likely to be related to (or to influence) changes in ridership in other routes.

More generally, it has been argued that segmentation of highway links can be based on any group of factors that significantly influence AADT. Garber and Bayat-Mokhtari [5], for example, found that AADT was significantly related to functional classification, functional use, land use, area population and terrain.

In regard to the extension of highway clustering methods to the transit environment, Deibel and Zumwalt [2] have identified the following factors as a possible basis for route clustering: functional classification (feeder, express, radial, cross-town), route length, headway,

number of stops, total boardings and peak load factor. Shanteau [18] found bus loads to be almost fully explained by headways, suggesting that this factor alone might serve as a basis for route clustering.

Empirical Analysis

In order to determine whether a set of factors could be identified to serve as a basis for evaluating the gains in precision associated with clusters of bus routes in the transit service network, we selected a sample of 17 routes with data recorded on the number of boarding riders during the April-June 1988 period. For each route we specified the following equation relating the number of boarding riders per bus trip to a set of time-of-service designations and a route configuration factor:

$$Ons = f(T_1, T_2, T_3, T_4, T_{1i}, T_{4o}, D_1T_1, D_1T_2, D_1T_3, D_1T_4, D_2T_1, D_2T_2, D_2T_3, D_2T_4, D_r), \text{ where}$$

T_1 = a dummy variable equalling 1 if the bus trip ended in the 7-9:00 AM period and 0 otherwise;

T_2 = a dummy variable equalling 1 if the bus trip ended in the 9-12:00 AM period and 0 otherwise;

T_3 = a dummy variable equalling 1 if the bus trip began in the 12-4:00 PM period and 0 otherwise;

T_4 = a dummy variable equalling 1 if the bus trip began in the 4-6:00 PM period and 0 otherwise;

T_{1i} = a dummy variable equalling 1 if the 7-9:00 AM bus trip was inbound and 0 otherwise;

T_{4o} = a dummy variable equalling 1 if the 4-6:00 PM bus trip was outbound and 0 otherwise;

D_1T_1 = an interaction term involving T_1 and a dummy variable equalling 1 if the bus trip was on a Saturday and 0 otherwise;

D_1T_2 = an interaction term involving T_2 and D_1;

D_1T_3 = an interaction term involving T_3 and D_1;

D_1T_4 = an interaction term involving T_4 and D_1;

D_2T_1 = an interaction term involving T_1 and a dummy variable equalling 1 if the bus trip was on a Sunday and 0 otherwise;

D_2T_2 = an interaction term involving T_2 and D_2;

D_2T_3 = an interaction term involving T_3 and D_2;

D_2T_4 = an interaction term involving T_4 and D_2;

D_r = a dummy variable equalling 1 for bus trips that deviate from the predominant origin-destination link defining the route.

The variables defined above are intended to capture the ridership differentials per bus trip in each of the selected routes on the basis of the time of day and day of the week the trip occurred, whether it was inbound or outbound during the weekday AM and PM peak commuting periods, and whether the trip involved a primary or secondary origin-destination. The disaggregation of time of service represented by the variables can also be considered a rough proxy for variation in both headways and passenger arrival rates by route.

A regression analysis was performed for each of the routes and the results are presented in Table 3. The coefficients associated with the variables in the equations are generally significant at the .05 level or better and the R-square values are moderately strong, indicating that the variables are capturing ridership variations related to time-of-day, direction, day of the week and origin-destination.

Table 3

Parameter Estimates for Selected "Determinants" of Bus Ridership by Route: April-June, 1988

	Route #																
Var.	4	6	8	9	12	20	33	43	54	57	70	71	72	75	108	109	120
α	23.4	14.8	14.2	17.3	19.2	11.2	24.9	14.7	15.0	35.4	10.2	22.2	22.6	26.2	10.1	27.2	16.2
t-val	14.1	15.9	11.9	15.5	9.9	15.9	7.7	5.8	10.8	6.0	7.1	11.4	14.7	17.2	7.3	18.0	22.8
T_1	8.4	4.0	0.4	1.4	9.6	3.6	7.1	-1.7	19.6	13.1	21.4	7.9	33.5	32.1	31.6	32.6	6.4
	1.9	1.5	0.1	0.5	2.1	1.8	0.8	-0.3	4.8	0.7	5.4	1.6	6.1	5.8	9.6	7.9	3.7
T_2	7.2	6.7	5.0	6.8	8.1	2.2	23.1	5.5	8.3	39.6	2.6	1.8	19.8	18.9	16.9	16.3	3.1
	2.3	3.2	2.1	3.1	2.5	1.5	2.9	1.4	3.0	3.1	1.0	0.6	5.0	4.8	6.4	5.4	2.3
T_3	23.8	17.1	20.0	12.5	18.5	7.2	27.1	10.9	17.7	51.9	12.8	16.2	45.8	43.7	19.4	33.4	10.8
	8.7	9.2	9.6	6.3	6.2	5.7	4.9	3.2	7.2	5.5	5.7	5.7	(3.6	12.7	8.0	12.3	9.5
T_4	12.0	17.4	15.9	3.9	21.5	4.7	17.6	9.5	15.3	28.9	9.8	13.7	46.9	40.9	18.5	19.3	7.4
	2.8	6.3	4.8	1.3	5.0	2.6	2.4	2.3	4.2	2.1	2.8	3.1	8.0	8.1	5.8	4.7	4.5
T_{11}	11.5	10.7	10.8	11.4	16.9	9.4	22.2	37.8	10.8	34.6	-8.6	11.3	-3.8	-2.1	-26	-6.0	-1.3
	2.0	3.6	3.3	3.2	3.3	4.0	2.2	7.0	2.2	1.8	-1.7	1.8	-0.7	-0.3	-6.8	-1.2	-0.6
T_{4o}	15.1	4.4	15.1	18.3	-8.9	7.6	23.9	2.9	8.1	37.2	-2.6	-4.8	2.1	-5.9	-2.6	19.5	-0.8
	2.8	1.5	4.4	5.5	-1.7	3.6	3.1	0.7	2.0	1.7	-0.6	-1.1	0.4	-1.0	-0.7	4.1	-0.4
D_1T_1		-13	-12	-7.9	-18	-14	-19		-19		-13	-14	-38	-30	-22	-24	-10
		-3.7	-3.2	-1.6	-2.1	-4.6	-1.8		-2.9		-1.1	-1.3	-5.8	-3.2	-4.4	-3.5	-3.9
D_1T_2		-4.6	-5.2	-2.9	-4.1	-0.6	-20		-0.1		-4.5	-0.1	-3.0	-7.0	-20	-13	-5.9
		-1.6	-1.5	-0.9	-0.6	-0.2	-2.1		-0.1		-1.1	-.01	-0.6	-1.1	-4.4	-2.9	-2.7
D_1T_3		-5.4	-6.3	-5.3	10.7	-1.2	-17		-1.2		-11	-18	-7.4	-5.3	-13	-25	-6.2
		-2.1	-2.2	-1.9	1.7	-0.6	-2.1		-0.3		-3.2	-5.1	-1.6	-1.0	-3.2	-6.7	-3.3
D_1T_4		-9.4	-10	-9.1	0.2	-3.4	-14		-17		-11	-12	-15	-12	-8.1	-17	-4.8
		-2.6	-2.5	-2.4	0.01	-1.3	-1.5		-2.6		-2.2	-2.3	-2.2	-1.6	-1.6	-3.4	-1.8
D_2T_1	-13	-11	-12	-11		-8.8			-29			-17	-39	-47	-26	-37	-3.2
	-1.4	-3.1	-2.6	-2.2		-2.4			-4.3			-2.2	-5.1	-6.4	-4.0	-5.3	-1.0
D_2T_2	-3.6	-9.0	-12	-3.4		0.1	-11	-4.4	-8.7	-33	-4.2	-9.5	-22	-20	-17	-13	-2.3
	-0.4	-3.1	-2.9	-0.8		0.1	-0.8	-1.0	-2.1	-1.5	-0.7	-2.3	-4.0	-3.5	-3.1	-2.2	-1.1
D_2T_3	-14	-13	-14	4.7		-1.1	-11	-5.6	-13	-13	-8.7	-19	-29	-27	-17	-17	-6.1
	-2.0	-4.9	-4.1	1.3		-0.5	-1.1	-1.5	-3.6	-0.8	-1.9	-5.4	-6.0	-5.6	-3.7	-3.6	-3.5
D_2T_4	1.1	-16	-16	1.8		-2.7	-24	-12	-7.9		0.8	-19	-34	-25	-11	4.6	1.3
	0.1	-4.5	-3.4	0.4		-1.0	-2.0	-2.4	-1.6		0.1	-3.6	-4.6	-3.8	-1.9	0.5	0.6
D_r	-3.3	-3.6	7.5	-6.7	-13	-6.3	-9.2	-2.2	-3.6	-10	5.6	-14	14.4	-6.8	2.8	-13	-12
	-0.7	-1.3	5.5	-2.1	-5.1	-2.8	-2.0	-0.6	-0.7	-1.4	1.1	-4.0	5.1	-1.2	1.3	-6.9	-16
R^2	.40	.48	.55	.39	.47	.32	.50	.70	.58	.49	.30	.50	.57	.56	.55	.56	.62
SEE	12.8	7.9	9.5	9.2	11.3	6.5	14.9	7.0	8.4	25.4	10.4	9.5	18.0	15.7	10.3	12.5	5.8
n	169	234	244	227	126	275	85	53	121	63	185	130	391	264	194	234	270

Examining the individual parameter estimates across routes, it is apparent that the consistency of the estimated changes in ridership varies by time of service. In particular, the estimates for both the AM peak inbound (T_{1i}) and PM peak outbound (T_{4o}) periods show considerable inter-route variation in magnitude. In some instances the values of the coefficients for these two variables are negative and statistically significant, indicating that counter-flow estimates of ridership during the peak commuting periods are actually higher. For example, the weekday AM peak period estimates of outbound ridership for routes 70, 72, 75, 108, 109 and 120 are higher than inbound ridership. These are cross-town routes, which would explain their apparent counter-intuitive ridership patterns. Other time periods with relatively high inter-route coefficient variation include Saturday morning and afternoon (D_1T_2, D_1T_3), and Sunday PM peak (D_2T_4).

High inter-route coefficient variation represents a serious challenge with respect to the application of clustering techniques to route level sampling and inference. It indicates that clusters of routes will be relatively less homogeneous at the bus trip level, thus limiting possible gains in precision. "Highway-type" methods of inferring ridership within clusters may thus hold less potential for application to the transit environment. A relatively more dissagregate sampling approach.may be needed to adequately reflect inter-route differences in ridership patterns, as a result.

The regression results provide a basis for static evaluation of the contribution of a priori knowledge about the effects of time of service on route sampling requirements. The variance in ridership for each of the routes sampled can now be defined in terms of one component explained by the regression and another representing unexplained ("error") variance. In the limit, we can utilize the information provided by the regressions to account for the explained variance, leaving the unexplained variance as the basis for determining route sample size requirements. This is illustrated in Table 4. The standard error of estimate from the regressions represents the

unexplained variation in boardings that is now substituted for the standard deviation term in the sample size equation presented earlier. The two right hand columns in Table 4 contain the sample sizes required to produce precision of +/- 20 per cent at the 90 per cent level of confidence for a conventional cluster sample, as determined by the total variation in boardings, and for the regression-based cluster sample, as determined by the unexplained variation component. Average sample size for the 17 routes with conventional sampling is 97 bus trips, versus an average of 53 bus trips per route with the regression-based sample. Application of the regression methodology thus results in a 45 per cent reduction in sampling requirements.

Table 4

Sample Sizes Required for a Conventional Cluster Sample and the Regression-Based Alternative

Rt. #	Mean	SD	SEE	n_{sd}	n_{see}
4	33.6	16.03	12.8	66.6	42.5
6	19.9	10.56	7.9	82.4	46.1
8	25.0	13.78	9.5	88.9	42.3
9	22.9	11.31	9.2	71.4	47.2
12	27.6	14.85	11.3	84.7	49.1
20	14.6	7.68	6.5	81.0	58.0
33	34.4	19.19	14.9	91.1	54.9
43	21.9	11.40	7.0	79.3	29.9
54	23.5	12.06	8.4	77.1	37.4
57	59.2	32.85	25.4	90.1	53.9
70	15.6	11.91	10.4	169.7	129.4
71	24.0	12.59	9.5	80.5	45.9
72	40.7	26.94	18.0	128.2	57.2
75	39.6	22.96	15.7	98.4	46.0
108	20.2	14.78	10.3	156.7	76.1
109	35.9	18.30	12.5	76.0	35.5
120	13.8	9.16	5.8	128.9	51.7
			Mean n	97.1	53.1

The reduction in sample size calculated for the regression-based approach represents an estimate in the limit, where the effects of time of service factors on ridership are fixed and known. In reality, however, these effects must be estimated from the sample data and thus cannot be known contemporaneously. We could assume that the effects of "time of service" on ridership are invariant from period to period, and use existing data to determine future sampling requirements. Such an assumption would be reasonable if the values of the various determinants of transit demand do not change greatly over time.

Analysis of Route Clusters

Given the results of the route-specific regressions, the next step in articulating a sampling plan involves segmenting the route network into homogeneous clusters. Cluster-specific regressions could then be estimated and used as a basis for estimating route level ridership from the sample data. The merits of this approach depend greatly on the extent to which variations in ridership patterns are minimized within clusters. For example, given a cluster comprised of n routes with identical ridership patterns, a sample of size N taken from within the cluster would yield the same precision as an alternative sampling plan employing nN observations and assuming independence between routes.

One means of evaluating the potential gains from route clustering involves determining the changes in the value of the coefficient of variation for the time of service regression parameters for progressive cluster sizes. With the 17 routes in our sample the number of possible clusters ranges from one, which would group all the routes together, to 17, which would treat each route as being independent. In the example above involving a single cluster of n identical routes, the associated coefficients of variation for the regression parameters would equal zero.

A factor must also be chosen as a basis for clustering. From those previously discussed we selected the number of boardings per bus trip. A cluster analysis was performed to create two, three and four group clusters of the 17 routes. Clustering was based on the objective of minimizing the composite within group variance of the average boardings per bus trip. The resulting allocation of routes to groups for the successive clusters is presented in Table 5.

Table 5

Composition of Route Clusters

Number of Groups								
2		3			4			
6	4	20	6	4	20	6	4	57
8	33	70	8	33	70	8	33	
9	57	120	9	57	120	9	72	
12	72		12	72		12	75	
20	75		43	75		43	109	
43	109		54	109		54		
54			71			71		
70			108			108		
71								
108								
120								

For each of the successive clustering arrangements the coefficients of variation (CV's) for the regression parameters reported in Table 3 were then calculated. These values are presented in Table 6. The issue of parameter instability discussed earlier is reflected in the CV values for the one-group clustering alternative; each of the parameters cited have CV values greater than one. Moreover, contrary to convention, the CV values for these parameters increase with the number of groups. Thus while the intra-group variation of average ridership decreases as the number of groups expands, intra-group variation in the AM peak inbound and PM peak outbound time slots actually increases. This result can be primarily traced to the definitional and operational differences associated with non-radial routes. Increases in CV values are also observed for Saturday PM (+9.2%), Sunday PM peak (+136.6%) and the "predominant route" variable (+12.4%). None of the parameters with declines in the CV value approach the percentage decline in intra-group average ridership variation.

Table 6

Coefficients of Variation for Ridership Regression Parameter Estimates, With Progressive Route Clustering

	Number of Group Clusters				Percentage
Parameter	1	2	3	4	Change (1-4)
α	.44	.22	.21	.18	-59.0
T_1	.94	.93	.95	.92	-2.1
T_2	.91	.63	.49	.41	-54.9
T_3	.63	.29	.25	.23	-63.5
T_4	.65	.48	.42	.41	-36.9
T_{1i}	1.89	1.96	15.60	15.86	+739.2
T_{4o}	1.64	1.97	2.14	2.12	+29.3
D_1T_1	.45	.29	.27	.27	-40.0
D_1T_2	1.01	1.05	1.00	1.00	-1.0
D_1T_3	1.09	1.17	1.19	1.19	+9.2
D_1T_4	.49	.45	.45	.45	-8.2
D_2T_1	.66	.54	.49	.49	-25.8
D_2T_2	.81	.68	.62	.56	-30.9
D_2T_3	.67	.65	.61	.59	-11.9
D_2T_4	1.12	1.03	2.72	2.65	+136.6
D_r	1.85	1.65	1.72	2.08	+12.4
Riders/trip	.42	.23	.15	.09	-78.6

1 The coefficients of variation for the 2, 3, and 4 group clustering alternatives reported in Table 6 are the weighted averages of the group-specific values of the statistic.

The near 80 per cent decline in the coefficient of variation for average riders per trip indicates that route clustering can result in a significant reduction in sampling efforts, so long as interest is limited to inferring total ridership at the route level. The gains from route clustering for more detailed inferences - by time period, direction and day of the week - are less noteworthy.

The latter conclusion is largely subject to the negative effects that the cross-town routes had on group homogeneity, however. These effects can be mitigated by several alternative modifications of the methodology employed above.

In developing the framework for route clustering, one possible modification would be to first segment routes into two general types - cross-town and "other" - and then proceed with clustering on the basis of ridership. This approach would distinguish two "populations" within the overall route network and would determine independent group clusters for each population. This approach could also be extended to include segmentation by weekday, Saturday and Sunday. A second approach would be to arbitrarily redefine (for statistical rather than route performance reporting purposes) the directional orientation of the cross-town routes, which would make their ridership patterns more consistent with the radial routes.

To the extent that the differences in ridership patterns between the cross-town and radial routes are limited to definitional issues the latter approach would be preferable because it would require fewer group clusters to decompose systematic ridership variance. Alternatively, if the cross-town and radial route ridership patterns differ in structure as well as definition, the former approach may be needed. Further analysis will be needed to identify which approach should be pursued.

Following the clustering of routes under the chosen segmentation alternative, the regression model (or some variant of it) could then be re-estimated for each of the route cluster groups. This would yield common ridership coefficients for the group members, as well as a common standard error of estimate which would be used to determine the necessary sample size for each group (as was done for each of the routes in Table 4). Also at this point, other more rigorous statistical techniques could be applied to evaluate whether successive route clustering produces a significant improvement in error variance (whose magnitude determines sample size) [13].

Conclusions

This report has explored possible ways of improving the precision of sample ridership estimates at the route level. Depending on the context of the application of the approaches developed in the report, one can conclude that the prospects for improvement in precision appear to be excellent, negligible or potentially fruitful.

The cluster analysis revealed that homogeneous groups of routes within the transit network can be identified on the basis of total boarding riders. The creation of four route groups reduced the coefficient of variation for average ridership from .42 to .09 (a 78.6 per cent decline). Very little inter-route variation remains within the groups, as a result, indicating that reasonable route level precision can be achieved with modest sample sizes. Based on previously analyzed data from the September-November 1988 period [20], we estimate that a route level sampling plan utilizing this approach would require an approximate 20 per cent increase in sample size over what is required at the system level. In contrast, a conventional sampling plan treating each route independently would require a near 300 per cent increase in the number of observations over the system level requirements. The drawback with the simple clustering approach is that ridership inferences are limited to route totals. While this change in scale represents an improvement over system level inferences, it may still be too general to provide a useful basis for route planning activities.

Coordination of the route clustering and regression approaches represents an attempt to recover ridership estimates at a scale that, for planning purposes, approaches the bus trip level. In regard to this exercise our findings are mixed; some of the time of service regression parameters cluster nearly as well as does route level ridership. But the gains for most of the parameters are slight, and some of the parameters actually exhibit greater variation after clustering. This latter finding is particularly troublesome because the coefficients involved include the weekday AM and PM peak periods, times of service for which special attention must

be devoted in scheduling and route planning. This approach, promising a small overall improvement in precision while yet lacking it for the more important trip segments, cannot be endorsed as it is presently defined.

Several possible modifications of the regression approach, involving the treatment of cross-town routes, promise improvements in precision for the trip segments in question. More generally, these modifications parallel refinements in the application of similar techniques to highway data collection [3]. A common issue in the transit and highway data collection arenas in regard to clustering is the initial specification of the systems' basic functional characteristics pertaining to the structure and level of use. Both the transit and highway data collection processes would benefit from the identification of a more representative typology defining the composition of the network. Improvements in this area would result in the identification of clusters that would produce more precise ridership inferences.

References

1. Chadda, H. and T. Mulinazzi, "A Transit Planning Methodology for Small Cities," Transit Journal, Vol. 3, No. 2, Spring 1977.
2. Deibel, L. and B. Zumwalt, Modular Approach to On-Board Automatic Data Collection Systems, Report 9, National Cooperative Transit Research & Development Program, Transportation Research Board, National Research Council, Washington, D.C., December 1984.
3. Dueker, K., R. Ledbetter and B. Rex, "Site Location Study for an Integrated Traffic Data Collection System," Phase I report to the Oregon Department of Transportation, Center for Urban Studies, Portland State University, July 1989.
4. Federal Highway Administration, U.S. Department of Transportation, Highway Performance Monitoring Field Manual for the Continuing Analytical and Statistical Data Base, 1984.
5. Garber, N. and F. Bayat-Mokhtari, "A Computerized Highway Link Classification System for Traffic Volume Counts," Transportation Research Record 1090, Transportation Research Board, National Research Council, Washington, D.C., 1985.
6. Gur, Y. and I. Hocherman, "Optimal Design of Traffic Counts," Paper presented at the 68th Annual Meeting, Transportation Research Board, Washington, D.C., January 22-26, 1989.
7. Horowitz, A., "The Design of a Single Route Ridership Forecasting Model," Transportation Research Record 980, Transportation Research Board, National Research Council, Washington, D.C., 1984.
8. Horowitz, A., "Simplifications for Single-Route Transit Ridership Forecasting Models," Transportation, Vol. 12, 1984.

9. Horowitz, A. and D. Metzger, "Implementation of Service Area Concepts in Single Route Ridership Forecasting," Transportation Research Record 1037, Transportation Research Board, National Research Council, Washington, D.C., 1985.

10. Horowitz, A. and D. Zlosel, "Transfer Penalties: Another Look at Transit Rider's Reluctance to Transfer," Transportation, Vol. 10, 1981.

11. Hsu, J. and V. Surti, "Framework of Route Selection in Bus Network Design," Transportation Research Record 546, Transportation Research Board, National Research Council, Washington, D.C., 1975.

12. Kyte, M. and J. Stoner, "A Time Series Analysis of Public Transit Ridership in Portland, Oregon, 1971-1982," Transportation Research-A, Vol. 22A, No. 5, 1988.

13. Maddala, G., Econometrics, New York, McGraw-Hill, 1977.

14. McClave, J. and P. Benson, Statistics for Business and Economics, 4th edition, San Francisco, Dellen Publishing Co., 1988.

15. Moser, C. and G. Kalton, Survey Methods in Social Investigation, New York, Basic Books, 1972.

16. Multisystems, Inc., An Assessment of Automatic Passenger Counters, Interim Report to the Urban Mass Transportation Administration, DOT-I-82-43, 1982.

17. Multisystems, Inc., Transit Data Collection Design Manual, prepared for the Urban Mass Transportation Administration Administration, DOT-I-85-38, June 1985.

18. Shanteau, R., "Estimating the Contribution of Various Factors to Variations in Bus Passenger Loads at a Point," Transportation Research Record 798, Transportation Research Board, National Research Council, Washington, D.C., 1981.

19. Shortreed, J., "Bus Transit Route Demand Model," Transportation Research Record 625, Transportation Research Board, National Research Council, Washington, D.C., 1977.

20. Strathman, J. and J. Hopper, "An Evaluation of Automatic Passenger Counters: Validation, Sampling and Statistical Inference," Discussion Paper 89-1, Center for Urban Studies, Portland State University, September 1989.

21. Urban Mass Transportation Administration, U.S. Department of Transportation, "Sampling Procedures for Obtaining Fixed Route Bus Operating Data Required Under the Section 15 Reporting System," Circular No. 2710.1, February 1978.