Use of Sampling in Bus-Line Data Collection

SUSAN P. PHIFER

Bus lines require periodic data collection for performance evaluation and efficient operation. The potential of using statistical methods in the collection of individual bus-line data is investigated. A particular bus line was chosen as an example. By using prior, conventionally collected data of that line, a suitable sampling method and estimation technique were developed. The sampling plan was applied to data collection in June 1980. The results demonstrate that sampling (a) gives the necessary precision while permitting more efficient use of manpower and (b) allows the flexibility needed to focus on data of interest. It is concluded that, with the establishment of appropriate guidelines for its use, sampling will provide a valuable tool to bus transit.

There is a need for regular evaluation of bus-line performance if a public transit system is to operate efficiently within a shifting environment. A valid evaluation process requires systematic data collection. Data-collection costs can be minimized through the use of statistical sampling methods. This paper describes an experience in adapting statistical methods to the realities of bus company operations and the encouraging results.

CURRENT SCRTD DATA-COLLECTION METHOD

Currently, the Southern California Rapid Transit District (SCRTD) collects bus-line data by making complete ride checks. All bus trips of a line are ridden by schedule checkers, and the boardings, alightings, and fares collected are noted for each bus stop. Schedule adherence information is taken at each time point along with any necessary explanatory notes that the schedule checker may see fit to make. As these data are collected for a line, they are processed by computer and used to update system-level data and create individual line reports. These line reports are used to plan service changes and tune schedules to optimize performance. This would appear to be an admirable amount of useful information, but there are three major factors that contribute to the inadequacy of this method:

1. Complete ride checks do not necessarily provide representative information. Many bus lines are too large to be ridden all in one day, so ride checks are scheduled over several days. If a checker should miss an assignment, it would usually not be completed until a later date. However, the people who use these data tend to use them as if they were collected on a single day. Furthermore, even a perfect and complete count of a single day's operation is only an approximation of the desired information. What is really needed is a forecast of data averages during the entire period for which a schedule will be in effect.
2. The complete ride check requires a considerable commitment of personnel. The work force of schedule checkers is insufficient to meet the objective of an annual ride check on each line, for weekday, Saturday, and Sunday schedules. The weekend schedules, especially, are frequently outdated since there are fewer weekend days on which to perform those checks.
3. The quantity of data collected in a complete ride check prohibits expedient processing of data to create the reports. This increases the lead time required before the collected data can be in the most readily usable form and precludes last-minute data collection.

Statistical sampling, although no panacea, can ameliorate the problems of representativeness and sufficient personnel. Instead of collecting complete line data, the statistical approach calls for strategically sampling representative data that can be used to make forecasts of data averages. The potential savings in checker hours from sampling can be applied toward a somewhat greater checking frequency, possibly enough to meet the once-per-year objective.

There are, of course, challenges to be dealt with on the way to implementation of statistical sampling. The most immediate is the need to establish methods appropriate to transit, since little statistical work has previously been done in this field.

SAMPLE SELECTION PLAN FOR BUS-LINE DATA

There are three areas to consider in the selection of a sampling plan: statistical, operational, and cost.

The selection of a statistically precise sampling plan requires a knowledge of the underlying characteristics of the data. A sampling plan that would collect data over the entire length of the route and, at times, spread throughout the day would ensure a more representative sample at various headways and passenger load levels. This would suggest using the individual bus trip as a sampling unit.

The sampling plan chosen must be amenable to practical assignment of the schedule checkers for actual data collection. These assignments must take into consideration the physical needs of the checker, the difficulty of creating the assignments, and the added costs associated with each.

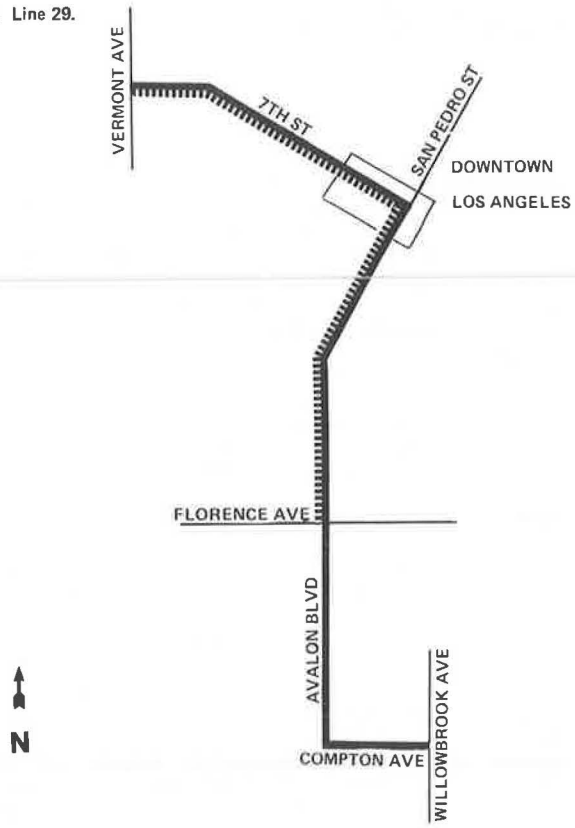
For these reasons, a simple random sample of individual bus trips could be wasteful of checker time and assignment time and therefore excessively costly. A more reasonable approach would be a cluster sample where individual bus runs of a line would be randomly selected and the bus trips within the bus runs would compose the sample. This plan would answer the statistical and operational needs while avoiding unnecessary cost.

BUS LINE SELECTED FOR CONTROL

A single bus line, line 29 on a Saturday schedule, was chosen for the initial development of the sampling process. Line 29 was chosen because there were data available from a complete ride check done in March 1980 that could be used to evaluate various sampling plans. In addition, line 29 had undergone a service change in June 1980, after the last data collection, and its effect had yet to be measured. The sampling plan derived with the March 1980 data would be applied to line 29 as an acid test of the operational component of the plan and to evaluate the effect of the June 1980 service change.

Line 29 is of moderate length, about 16 miles, running north from the City of Compton to downtown Los Angeles and then west along Seventh Street to Vermont Avenue. On the south leg the line has a short turn at Florence Avenue and Avalon Boulevard, about halfway between downtown Los Angeles and the Compton terminus (see Figure 1). Operation is divided approximately equally between the full route and the short line. An aggregated headway of 10 min is maintained, on the northern portion of the line, throughout most of the day. Line 29 has a generally

Figure 1. Line 29.



high level of ridership and is one of the more heavily traveled lines on Saturdays.

The Saturday schedule of line 29 was chosen in order to develop sampling techniques appropriate for weekend schedules. Sampling techniques are urgently needed for weekend data collection because of the limited number of checker days available.

SUITABLE ESTIMATION TECHNIQUE FOR RIDERSHIP

This work develops a technique for estimating bus-line ridership by using sampled data. Ridership estimation is an initial but critically important step toward the eventual development of a comprehensive estimation scheme, including schedule adherence and fare breakdowns, that can frequently supplant the need for a full ride check.

A good ridership estimation method should be sensitive to the high degree of variability from one bus trip to the next. It has been demonstrated that passenger loads vary mainly as a result of buses bunching (1). A late bus could carry a disproportionately heavy load and the bus following it the reverse. To account for the bunching effect, some information from every trip should be used as a regression estimator of boardings per trip. Using information from every trip would also serve to correct for bias introduced by the presence of checkers on the sampled bus runs.

A variety of information could be collected from every trip. Several variables could conceivably provide adequate estimates of boardings. Criteria for variable selection should include desired precision and reasonable cost of data collection. Possible variables could be boardings, alightings, passengers on board, and combinations of these.

It was initially presumed that data from each trip would be collected by means of a standing point check at one bus stop. An active bus stop on line 29 (Alvarado and Seventh Streets) was selected. The variables listed above were each tested for their degree of correlation with passengers per bus trip by using the March 1980 data. Results of the analysis showed that boardings and alightings combined provided the highest correlation. The form of the regression equation with combined boardings and alightings as the independent variable is as follows: Estimated boardings per trip = B (boardings + alightings) + constant, where the boardings and alightings are data collected at a bus stop and B and the constant are derived by using standard regression formulas and the boardings per trip from the sample of complete trips.

However, information gathered for a single bus stop contains a significant component of random fluctuation unrelated to the trip ridership. Estimates of passengers per trip must be accurate since decisions that will affect daily line operating costs are based on them. Therefore, the potential for collecting data on boardings and alightings over a small segment of line 29 rather than at one particular stop was investigated. Collecting data on a line segment also allows an opportunity to focus on a section of particular interest. In the case of line 29 this would be the western section, where specific effects of the service change could possibly be observed.

Data collection over a segment seems logical because it is usually cheaper to have checkers ride a segment than to conduct point checks at each bus stop along the segment. It was hypothesized that as the length of the segment increased so would the precision of the estimate. In addition, cost depends on the number of schedule checkers needed to cover a certain length segment. This does not increase at regular distance intervals but is affected by more complex factors related to the scheduled running time of the bus over the segment and variability in actual running time.

To evaluate relative precision and cost, three segments of different lengths, each containing the western section of line 29, and the point check mentioned previously were analyzed. Figure 2 shows the point check and the three segments and the relative projected costs of their data collections (figured in checker hours). The column labeled "aggregate" in Table 1 gives the correlations of total boardings per trip with the point-check variable (combined boardings and alightings) and with each of the three segment variables (sum of total boardings and alightings throughout the segment).

As can be seen, the longest segment had a 20 percent improvement in correlation over the point check. The added one-time cost of about 50 checker hours would bring a significant gain in precision, allowing dependable passenger estimates while still saving 50 percent in data-collection costs over the full riding check.

SUMMARY OF RIDERSHIP ESTIMATION PLAN

To review, the use of boardings and alightings over a short portion of a line provided a better regression estimate of passengers per trip than the use of one stop alone. Estimate precision was further improved by stratifying the line data by direction of travel and by subroute (full route or short line). Conditions are more similar within these groups. Table 1 indicates the improvement in the correlation of the variable when the data are broken down into the above strata. In using stratification, a separate regression equation would be derived for each combination of subroute and trip direction. Estimates of total passengers within each stratum would be calculated, and then they would be combined to provide overall totals.

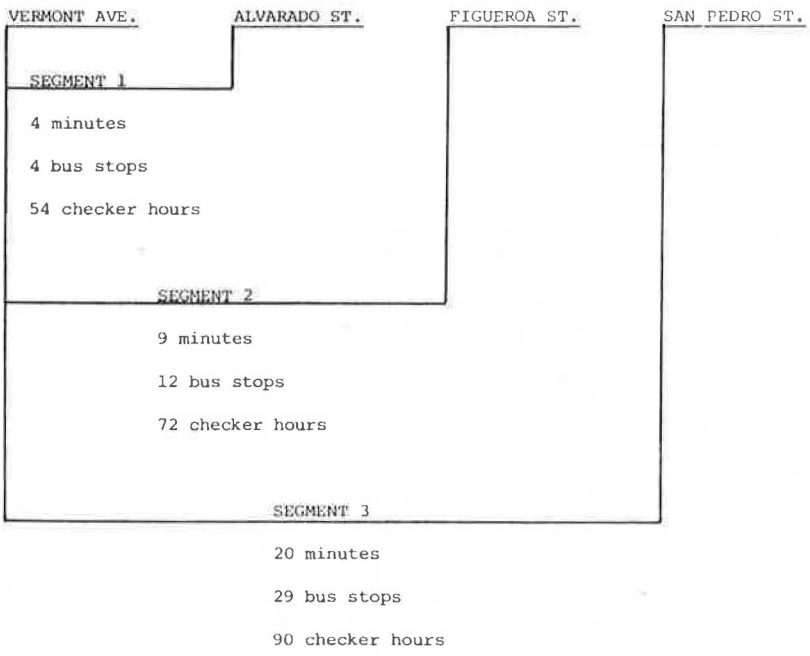


Figure 2. Point check and segments analyzed.

Table 1. Comparison of variable correlation with boardings per trip for stratified and unstratified data.

	Correlation Coefficient				
		Eastbound		Westbound	
Variable[a]	Aggregate	Short Line	Full Route	Short Line	Full Route
Point check	0.7448	0.8099	0.8147	0.8226	0.7752
Segment 1	0.7342	0.8863	0.7610	0.8457	0.8174
Segment 2	0.7996	0.8913	0.7904	0.9095	0.8340
Segment 3	0.8901	0.9712	0.9174	0.9721	0.9007

[a]Boardings + alightings.

Table 2. Precision of estimated boardings per trip at various sample sizes.

Sample Size (%)	Direction	90% Confidence Interval (no. of boardings)	Portion of Total Boardings[a] (%)
15	Eastbound	±1520	±13
	Westbound	±487	±4
	Total	±2007	±8.5
30	Eastbound	±578	±5
	Westbound	±341	±3
	Total	±919	±4
45	Eastbound	±453	±4
	Westbound	±298	±2.5
	Total	±751	±3

[a]Total boardings = 23 437.

SELECTING A SAMPLE SIZE

In addition to selecting an appropriate sampling plan and developing an estimation method, it was necessary to select an appropriate sample size. Three samples of different sizes were chosen from the March 1980 data. The samples consisted of approximately 15, 30, and 45 percent of the total bus trips. These were chosen by using the sampling plan selected earlier and randomly sampling the bus runs until the number of bus trips within the runs approximated the desired sample size. The ridership estimate was obtained for each sample size, and the respective 90 percent confidence intervals were calculated for comparison. Table 2 displays these results.

As can be seen, the confidence interval for the 15 percent sample size is quite large, especially for the eastbound direction of line 29, which has demonstrated more variability in passengers per bus trip. A ridership estimate with this much leeway would be of little value in determining whether buses on the line were overcrowded, for instance, since this confidence interval would allow ±13 percent passengers per bus trip. However, the confidence interval for the 30 percent sample gives ±5 percent precision, which would be considerably more reliable. The gain in precision with the 45 percent sample is slight. Use of a 30 percent sample would be optimum in this case.

The complete sampling plan presented here, when applied to line 29, saves the use of 10 schedule checkers. This represents a 44 percent savings in cost from a full ride check, which does not include the potential savings in processing sample data.

APPLICATION OF METHOD

The plan as developed to this point theoretically meets the requirements of bus-line data collection. However, only in actual use can comprehensive evaluation take place. Preparation was made for sample data collection on line 29.

The creation of the schedule checker assignments

Table 3. Estimates and 90 percent confidence intervals from March 1980 data.

Direction	Estimated Boardings ± 90 Percent Confidence Interval			Actual Boardings		
	Short Line	Full Route	Total	Short Line	Full Route	Total
Westbound	4300 ± 135	5587 ± 282	9 887 ± 417	4152	5638	9 790
Eastbound	4558 ± 119	5581 ± 257	10 139 ± 376	4797	5628	10 425
Total			20 026 ± 793			20 215

required more effort than for a full ride check, as had been expected. The checker assignments that covered the segment portion of the line were of necessity inefficient. Past experience in assigning schedule checkers has shown that, when they must make transfers, even a liberal assignment schedule frequently leads to a missed connection. A prior knowledge of the expected variability of schedule adherence for the line being assigned can be important. In the case of line 29, an assignment schedule that at times allowed a schedule checker a half-hour or more between bus rides was necessary to ensure that no trips were missed. This was actually somewhat generous, but the alternative of a rotation with one less checker was too restrictive to allow for normal delays.

Due to the hazards of nighttime transfers, it was decided to suspend the segment ride check at 7:00 p.m. The choice existed to continue the check as a full check for information on the night ridership, but for the purposes of this data collection this information was felt to be unnecessary. In any case, the efficiency of sampling is lost when headways lengthen and only a few bus runs remain in operation.

RESULTS OF JUNE 1981 DATA COLLECTION

The results of the June 1981 data collection are presented below:

Direction	Estimated Boardings ± 90 Percent Confidence Interval		
	Short Line	Full Route	Total
Westbound	2386 ± 114	2976 ± 215	5 362 ± 329
Eastbound	2544 ± 135	3509 ± 204	6.053 ± 339
Total			11 415 ± 668

Table 3 gives parallel estimates from the March 1980 control data. These were computed by using an identical sampling plan. The 90 percent confidence interval for the March 1980 estimate allows a ±4 percent error. As can be seen in Table 3, the actual number of passengers carried is within 1 percent of the estimated number. For the June 1981 sample data collection, the 90 percent confidence interval permits a ±6 percent error. These results do not contradict the hypothesis that the variance used in these calculations in fact overestimates the error.

In comparing the March 1980 with the June 1981 ridership figures, it is immediately apparent that there was a large decline in ridership (43 percent). There are at least two reasons for this difference. The main reason, as mentioned previously, is that service changes that were expected to reduce ridership occurred on line 29 in June 1980 (after the March 1980 data collection). The seasonal variation in system patronage is another contributing factor.

COMMENTS AND CONCLUSIONS

As the analysis of sample data collection continues, a number of additional issues must be explored. Estimation of fares collected and schedule adherence need investigation. In addition, more experience is required with sampling plans to allow refinement of the variance estimate. As sampling develops, complementary data processing methods must be established and users will require education in the correct use of the data.

It is important to evaluate sample data collection in light of its strengths and weaknesses in order to use the method optimally. It is recognized that additional costs are incurred by the extra time required to write schedule checker assignments and by the less-efficient use of checker time. The extent of these additional costs, both now and over the long run, remains to be established. In addition, possible savings to be gained with sample data collection diminish as headways lengthen, and at some point this results in a complete ride check being more economical.

However, one great advantage of sample data collection is its flexibility. A sampling plan can be varied to gain quantity and accuracy of information where they are most desired and thus save on data-collection costs. By applying the savings to obtain more frequent data collection, seasonal variations can be estimated and important trends can be recognized. It seems very likely that sample data collection will develop into an efficient tool for bus-line data collection.

ACKNOWLEDGMENT

Joel Woodhull provided overall guidance on this paper. Anne Huck, Ed Vandeventer, and the SCRTD Schedule Checking Section gave their invaluable assistance and constructive comments. The extensive review by Jesse Simon is greatly appreciated. Special thanks to Susan Correa for her patient assistance in preparing the manuscript.

REFERENCE

1. R.M. Shanteau. Estimating the Contribution of Various Factors to Variations in Bus Passenger Loads at a Point. TRB, Transportation Research Record 798, 1981, pp. 8-11.

Publication of this paper sponsored by Committee on Public Transportation Planning and Development.