PROCEEDINGS —

Fifteenth Annual Meeting

Theme:
"Transportation in Focus"

October 10-11-12, 1974
Fairmont Hotel
San Francisco, California

Volume XV • Number 1 1974

TRANSPORTATION RESEARCH FORUM

THE TRAVEL forecasting procedure developed primarily in connection with urban transportation studies is one of the most widely used macro-simulation techniques in transportation planning. The procedure, consisting of the three steps of trip generation, trip distribution and traffic assignment, has been standardized to a large extent in the course of its numerous applications of the procedure, its results, however, are not accepted without questioning (1, 2, 3). Transportation planners actually are concerned about some of the limitations of the procedure and are constantly striving to improve its reliability. Many investigations have been made to refine the techniques used at each of the steps of the procedure for better performance. However, there are certain basic constraints that limit the reliability of the results obtained from this set of travel forecasting models. The most important of such limiting factors is the reliability of the travel data that are used to develop and calibrate these models.

The travel data that are used to develop the simulation models for traffic forecasting are obtained primarily from the origin-destination (O-D) surveys, which are generally of two types—home-interview and roadside-interview survey. The home-interview survey data are used to develop the existing internal travel pattern, whereas the roadside-interview survey data are used for existing external-internal and through travel. While the O-D survey data are necessary for deriving the mathematical models for trip generation and trip distribution, the traffic assignment models are based on theoretical hypotheses and do not use O-D information. The mathematical simulation of travel is accomplished by the sequential use of these models. In developing the existing travel patterns, however, the zone to zone distribution of trips can be derived directly from the O-D survey data, thus omitting the step of trip generation.

The reliability of travel data is directly related to the sample size, which is the main subject of this investigation. The sample rates recommended by the Federal Highway Administration (4) for home-interview surveys in cities of different sizes are presented in Table 1. These rates are based primarily on the trip making characteristics and their variations at the household level. The adequacy of the recommended sample size with respect to the reliability of the derived travel pattern either at the distribution or traffic assignment stages is not explicitly considered in the sampling design. However, in the actual planning process, the transportation planners are mostly concerned with the reliability at the distribution and/or assignment stages. In the following sections of the paper, the sample size requirements for specific levels of reliability at both the distribution and assignment stages are examined. The analytical approach to be presented in this paper is basically applicable to both internal and external trips. However, the discussion and the actual analysis is primarily oriented towards internal travel and the sampling procedure for home-interview surveys.[1]

FHWA Recommended Sample Rates for Home-Interview Origin-Destination Survey

Population of Area	Recommended Size of Sample
Under 50,000	1 in 5 dwelling units
50,000 to 150,000	1 in 8 dwelling units
150,000 to 300,000	1 in 10 dwelling units
300,000 to 500,000	1 in 15 dwelling units
500,000 to 1,000,000	1 in 20 dwelling units
Over 1,000,000	1 in 25 dwelling units

TABLE 1

RELIABILITY AT TRAFFIC ASSIGNMENT STAGE

A common way of verifying the simulated travel is to compare the assigned link volumes with actual ground-counts. The accuracy and reliability of the traffic volume estimates on each link can be examined on either an individual or simultaneous basis. Vaughan (5) investigated the reliability at the traffic assignment stage following the concept of 'individual link reliability' and his approach, a pioneering effort, will be discussed first before the introduction of other concepts.

Vaughan's Approach

Vaughan (5) analyzed the reliability of individual link volumes using a spider network in which the centroids representing the traffic zones are the only nodes being inter-connected with each other either directly or through other nodes. A spider network is actually an over-simplification of a typical network used in any urban transportation study. Again, the traffic assignment technique used by Vaughan was also very simple compared with some of the sophisticated techniques, such as the capacity restrained and stochastic assignments. Actually in assigning zone to zone travel on the spider network, he used predetermined proportions to allocate a particular movement among two alternative routes on the basis of their difference in

Reliability Analysis of Origin-Destination Surveys and Determination of Optimal Sample Size

by Gary G. Makowski; Arun Chatterjee**; and Kumares C. Sinha****

travel impedance. The analysis performed by Vaughan is presented in this section in a summary form. The readers interested in the detailed analysis should refer to his original paper (5). The notations and the data of the numerical example used in Vaughan's paper are used also in the subsequent sections of this paper in order to maintain consistency and continuity.

Vaughan's Notations and Assumptions: **In the analysis of home-based work trips, Vaughan used the following notations and assumptions:**

h_i = number of commuters of home zone i;

π_{ij} = the proportion of commuters of home zone i who work in work zone j, ($\Sigma_j \pi_{ij} = 1$ and $\pi_{ii} = 0$, for all i)

$\tilde{\pi}_{ij}$ = an estimate of π_{ij};

T_{ij} = number of commuter trips from home zone i to work zone j;

$\tilde{T}_{ij}$ = an estimate of T_{ij}, ($\tilde{T}_{ij} = h_i \tilde{\pi}_{ij}$);

$a_{ij}(kl)$ (footnote [2]) = the proportion of traffic from zone i to zone j that uses the link (k, l);

μ_{kl} = average traffic flow on link (k,l);

$\tilde{\mu}_{kl}$ = an estimate of μ_{kl}

If n_i is the number of commuters sampled out of a total number of commuters h_i in zone i, then the estimate $\tilde{\pi}_{ij}$ of π_{ij} is the number of commuters, say, e_{ij}, in the sample who work in j divided by n_i. Thus,

$$\tilde{\pi}_{ij} = e_{ij}/n_i$$

and $$\tilde{T}_{ij} = h_i \tilde{\pi}_{ij}$$

Vaughan assumed $\tilde{T}_{ij}$ to be binomially distributed with the mean $h_i \pi_{ij}$ and variance $h_i \pi_{ij}(1 - \pi_{ij})$. Using the normal approximation to the binomial, Vaughan deducted that

$$\tilde{T}_{ij} \sim N\,[h_i \pi_{ij}, h_i \pi_{ij}(1 - \pi_{ij})]$$

$\tilde{T}_{ij}$ is actually the estimated value at the distribution stage of the travel simulation procedure and its reliability is examined in a later section. The value of interest at the assignment stage is μ_{kl}, which may be expressed as,

$$\mu_{kl} = \Sigma_{ij}\, a_{ij}(kl)\, E\,(\tilde{T}_{ij}) = \Sigma_{ij}\, a_{ij}(kl)\, h_i \pi_{ij}.$$

The estimator $\tilde{\mu}_{kl}$ of the traffic flow μ_{kl} is then taken to be

$$\tilde{\mu}_{kl} = \Sigma_{ij}\, a_{ij}(kl)\, h_i \tilde{\pi}_{ij},$$

so that $\tilde{\mu}_{kl}$ has a mean or expected value of μ_{kl} and a variance V_{kl} approximated by

$$\Sigma_{ij}\, a_{ij}^2(kl)\, h_i^2 \pi_{ij}(1 - \pi_{ij})/n_i.$$

Using $s_{ij}^2 = h_i^2 \pi_{ij}(1 - \pi_{ij})$, V_{kl} can be written as

$$V_{kl} = \Sigma_{ij}\, a_{ij}^2(kl)\, s_{ij}^2/n_i$$

Since $\tilde{T}_{ij}$ has approximate normality, so does $\tilde{\mu}_{kl}$ and thus,

$$\tilde{\mu}_{kl} \sim N(\mu_{kl}, V_{kl}).$$

Vaughan's Optimal Sample Size: **Vaughan used two different approaches to derive optimal sample sizes. The objective of the first approach was to minimize cost to achieve a given level of accuracy, while the other approach had the objective of minimizing the error for a given financial budget for sampling. This study is concerned with the first approach where Vaughan required that the μ_{kl} to be within δ of its true value with probability $(1 - \alpha)$ and at a mini-**

**Department of Mathematics and Statistics, Marquette University, Milwaukee, Wisconsin*

***Department of Civil Engineering, Marquette University, Milwaukee, Wis.*

****Department of Civil Engineering, Marquette University, Milwaukee, Wis.*

mum cost. Vaughan's formula for optimal sample size n_i is,

$$n_i = \max_{kl} [n_i(kl)] \quad \ldots\ldots\ldots (1)$$

where $n_i(kl)$ is the optimal sample size in zone i considering just one link (k, l) alone, and is expressed as

$$n_i(kl) = \frac{z_i(kl) \sum_i c_i z_i(kl)}{V_{kl}} \quad \ldots\ldots (2)$$

where $z_i(kl) = [\frac{1}{c_i} \sum_j a_{ij}^2(kl)\, s_{ij}^2]^{1/2}$, c_i is the cost of sampling in the ith home district and letting $V_{kl} = \frac{\delta^2}{U_\alpha^2}$, U_α being the two sided outer α percent cut-off of the standardized normal distribution. Thus $U_{.05} = 1.96$)

Vaughan demonstrated his technique of determining optimal sample size with a numerical example. For a small city of 140,000 commuters, he used seven districts (or zones) and a spider network. The thirteen links of the network and their travel impedance are shown in Figure 1. The population and employment in each zone is given in Table 2. Vaughan estimated the values for $\tilde{\pi}_{ij}$ using the concept of a 'gravity model' (2), that is,

$$\tilde{\pi}_{ij} = \frac{w_j f_{ij}^{-1}}{\sum_j w_j f_{ij}^{-1}}$$

where w_j = number of employees in district j, and f_{ij}^{-1} = friction factor based on travel impedance between zones i and j.

The value of $a_{ij}(kl)$ was estimated for

SPIDER NETWORK USED FOR VAUGHAN'S NUMERICAL EXAMPLE

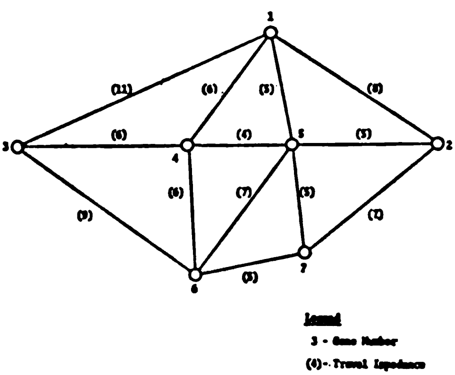

FIGURE 1

Commuter Population and Employment in Each Zone

Zone	Commuter Population (h_i)	Employment (w_j)
1	13,680	8,720
2	14,000	25,590
3	540	11,930
4	12,470	39,870
5	29,230	29,770
6	33,710	13,120
7	36,370	11,000
Total	140,000	140,000

TABLE 2

each link of the network on the following basis:

For each pair of zones, two alternative paths are determined on the basis of the 'minimum cost' criterion. If the travel impedances on the two paths are equal, traffic is assigned on both routes on a 50 per cent-50 per cent basis; if the costs differ by one, traffic is assigned on a 70 per cent-30 per cent basis; if differing by two, a 80 per cent-20 per cent basis is used; if differing by three, a 90 per cent-10 per cent is used; and if differing by four or more, all traffic is assigned to the 'minimum' path.

For the purpose of this study, Vaughan's procedure and the numerical example was used to compute the optimal sample size on an individual link basis. The computations were based on $\alpha = 0.05$, $\delta = 700$[8] and $c_i = 1$. The values of $\tilde{\pi}_{ij}$ and $a_{ij}(kl)$ computed by Vaughan were used in this analysis; but they are not included in this report to avoid unnecessary duplication. The derived sample size for each zone is shown in Table 3.

Limitations of Vaughan's Approach

Vaughan's approach, based on theoret-

SAMPLES BASED ON INDIVIDUAL LINK RELIABILITY
(for Vaughan's Home to Work Trips)

	Sample Size and Rate 95 Per Cent Individual Reliability	
Zone	Sample Size (Commuters)	Rate[a] in Per Cent
1	1,581	11.6
2	2,571	18.4
3	104	19.3
4	2,938	23.6
5	5,853	20.0
6	3,447	10.2
7	6,870	18.9
Total	23,364	16.7

TABLE 3

ical analysis, has a few practical limitations. The scope of his analysis is also limited since he includes only home to work trips that constitute less than 25 per cent of all trips made by the households. It must not be overlooked that the reliability of Vaughan's approach refers only to that of the home to work trips and not the total traffic volume.

Vaughan's analysis of home to work trips was based on the number of commuters in each zone and the respective proportions of commuters residing in a zone who work in another given zone. Thus the sample sizes based on Vaughan's procedure are those of commuters and it must be noted that if the normal procedure of household survey is followed, the sample of all persons that would be necessary to ensure the selection of the number of commuters required by Vaughan's estimates, will be larger.

Regarding appropriate measures of travel demand, it must be pointed out that the unit, 'vehicle per/hour', used by Vaughan to express the allowable error, $\delta\,(=700)$, is not compatible with other assumptions. Vaughan overlooked the significance of 'car occupancy' factors that are necessary to convert person trips to vehicle trips.

One significant limitation of Vaughan's numerical example that must be pointed out is related to the small number of zones. An examination of equation 2 reveals that the sample size is directly proportional to $(h_i)^2$, where h_i is the population of zone i. The number of zones, which is related to zonal population, has a significant effect on the overall sample size. The use of only nine zones for an urban area of 140,000 commuters is far from realistic and affects the overall sample size seriously. Considering the inclusion of only home to work trips and the use of nine zones in determining the sample size, the results of Vaughan's procedure can not be compared with what is obtained by using FHWA guidelines.

In addition to the above limitations of some of the procedural assumptions, the practical usefulness of Vaughan's approach is questionable too. The estimation of the values of π_{ij} and $a_{ij}(kl)$ in the case of typical urban transportation studies would require a significant amount of effort and cost. Although the estimation π_{ij}'s is necessary for the gravity models used for trip distribution, they are generally derived at a later stage of the study. Derivation of π_{ij} values at the beginning of the study would imply the use of synthetically developed gravity models. The estimation of the values of $a_{ij}(kl)$ at the beginning of a study is even more difficult and almost impossible in the case of capacity restrained or stochastic assignment techniques. Vaughan's treatment of $a_{ij}(kl)$ as a deterministic variable as opposed to a stochastic variable is also questionable. Considering the cost and trouble of estimating the values of π_{ij} and $a_{ij}(kl)$ prior to the actual phase of model calibration, Vaughan's procedure is impractical.

The sampling procedure developed by Vaughan provides a measure of the reliability of the estimated traffic volumes on each link considered separately, that is, individually. Such a measure of reliability does not describe directly the overall or simultaneous reliability that is commonly used in the analysis of a system, such as a transportation network. The concept of simultaneous reliability and its effect on sample size are discussed in the following section.

Simultaneous Link Reliability

The concept of simultaneous reliability may be explained with the well-known example of 'light bulbs'. For instance, the reliability of each light bulb in a group of, say, fifty, may be analyzed on an individual basis and it may be concluded that the probability that any bulb selected from the group will burn for one month without failure is 90 per cent. However, the probability of all the fifty bulbs burning simultaneously for one month without failure is likely to be much less than 90 per cent. The reliability in the first case is on an individual basis, whereas the latter is a case of simultaneous reliability.

In the case of a transportation network it is quite natural to ask about the probability of obtaining reliable estimates on all the links of a network simultaneously. Vaughan's measure of reliability does not provide an answer in that direction. For instance, in the previous numerical example, there was 95 per cent probability that the error on each link will be less than 700; however, the probability that such an accuracy will be obtained on all the links simultaneously is likely to be much less. In other words, the chances of making an error of 700 on at least one of the thirteen links is larger. To make a rough guess about this overall reliability, the errors on each link may be assumed to be independent random variables (which they are not), and then the occurrence of errors at least as large as 700 (i.e. $\geqslant 700$) would itself be a binomial random variable. Thus with reference to the example problem with 13 links, the number of trials would be 13 and the probability of a large error ($\geqslant 700$) on each link (or trial) is .05. The probability of at least one large error ($\geqslant 700$), therefore, would be $1-(.95)^{13} = .49$.

The above crude analysis of overall reliability clearly reveals the significant difference between the two concepts of 'individual reliability' and 'simultaneous reliability'. The sample size requirement for obtaining simultaneous reliability at the same level as the individual reliability, is also expected to be significantly different and will be explored in this section of the report. In addition, simultaneous confidence intervals will also be derived for the estimated link volumes.

Bonferonni Procedure for Optimal Sample Size: Bonferonni procedure (6) is basically a technique to combine individual reliabilities together. Using the property of subadditivity of probabilities,

$$P(\text{Error} > \delta \text{ on at least one link}) \leqslant \sum_{\ell=1}^{N} P(\text{Error} > \delta \text{ on link } \ell),$$

where N = number of links in a network.

Thus to ensure a simultaneous reliability of $(1-\alpha)$, the individual link reliability must be achieved at the level $(1-\alpha/N)$.

With reference to the numerical example being used in this paper, for 95 per cent individual link reliability, the value of V_{kl} was computed as follows:

$V_{kl} = \delta^2/U^2_{.05} = (700)^2/(1.96)^2 = 127{,}551.02$,

where 1.96 is the two sided outer 5 per cent cut-off of the standardized normal distribution. However, to ensure an overall reliability of 0.95 for all links simultaneously, the error probability of α, which is .05 in this case, must be divided by the number of links, which is 13 in this case. Thus, the value of V_{kl} to be used for simultaneous reliability is given by the formula,

$$V_{kl} = \delta^2/U^2_{\alpha/N} \quad \ldots\ldots\ldots\ldots \quad (3)$$

and for the numerical example,

$\alpha/N = .05/13 = .003846$ and $V_{kl} = (700)^2/U^2_{.003846} = (700)^2/(2.89)^2 = 58667.88$

Using this new value of V_{kl} in the previous equations 1 and 2, new sample sizes were obtained, which are presented in Table 4. The sample size obtained by this procedure ensures an overall reliability of 0.95, that is, there is 95 per cent probability that the error in estimating the travel demand on one or every link will not be greater than 700. As the comparison of Tables 3 and 4 will reveal, the ratio of the sample size to ensure 95 per cent simultaneous reliability on all links and that necessary to ensure 95 per cent accuracy on each link separately, is equal to the ratio of the two V-values used for the respective approaches, which is 127,551.02/58,667.88 = 2.17.

SAMPLES BASED ON SIMULTANEOUS LINK RELIABILITY

(for Vaughan's Home to Work Trips)

Sample Size and Rate 95 Per Cent Simultaneous Reliability

Zone	Sample Size (Commuters)	Rate in Per Cent
1	3,436	25.1
2	5,589	39.9
3	225	41.7
4	6,386	51.2
5	12,724	43.5
6	7,494	22.2
7	14,934	41.1
Total	50,788	36.3

TABLE 4

Simultaneous Confidence Intervals: A confidence interval for each of the link volumes can be derived on the simultaneous basis using the Bonferonni approach. The simultaneous confidence interval can be estimated using the following formula:

$$\Sigma_{ij}\, a_{ij}(kl)\, h_i\, \tilde{\pi}_{ij} - U_{\alpha/N}\, \tilde{V}_{kl}^{1/2} \leqslant \mu_{kl} \leqslant \Sigma_{ij}\, a_{ij}(kl)\, h_i\, \tilde{\pi}_{ij} + U_{\alpha/N}\tilde{V}_{kl}^{1/2} \quad \ldots \quad (4)$$

where $\tilde{V}_{kl} = \Sigma_{ij}\, a_{ij}^2(kl)\, h_i^2\, \tilde{\pi}_{ij} (1-\tilde{\pi}_{ij})/n_i$

and other notations are as described before. The procedure outlined in equation (4) assures that "$(1-\alpha)$ per cent of the time" the confidence intervals for all the links will be valid simultaneously, that is they will contain the actual expected link flow.

As an illustration of the procedure, the numerical example of this paper is used to obtain 95 per cent simultaneous confidence interval for the expected link flow of the link (3.6). With $\alpha = .05$ and N = 13,

$U_{\alpha/N} = U_{.05/13} = U_{.003846} = 2.89$

Using the sample sizes given in Table 4,

$3660.19 - (2.89)\,(102.69) \leqslant \mu_{36} \leqslant 3660.19 + (2.89)\,(102.69)$

or $3363.42 \leqslant \mu_{36} \leqslant 3956.96$

The above interval for the traffic volume on link (3, 6) along with those for the other twelve link flows will simultaneously enjoy 95 per cent confidence.

RELIABILITY AT TRAFFIC DISTRIBUTION STAGE

The zone to zone distribution of travel within an urban area has special significance in transportation planning. Although the travel demand in the form of zone to zone distribution is not related to specific routes, it represents the basic desire of travel and one of the important objectives of an origin-destination survey is to develop the non-route specific desire lines. Actually the traffic volumes on specific routes are relatively unstable in the sense that alternative routes may be used to satisfy a specific zone to zone travel desire. On the other hand, for certain trip purposes, such as home to work trips, the origin and destination of most trip-makers are fixed. Consequently, the zone to zone travel estimates are considered the fundamental basis for transportation planning. The origin-destination survey is directly related to the zone to zone distribution of travel, which is actually derived by expanding the sample survey by appropriate factors.

Modifying the previous definition of π_{ij}, the proportion of commuters in home zone i who work in zone j, and **redefining it to be the proportion of home to work trips that are produced in zone i and attracted to zone j**,[4] the value of the π_{ij} is estimated by $\tilde{\pi}_{ij}$, the relative frequency of home to work trips reported in the O-D survey that are produced in zone i and attracted to zone j. The value of T_{ij}, home to work trips from home zone i to work zone j, is estimated as follows:

$$\tilde{T}_{ij} = P_i \tilde{\pi}_{ij},$$

where P_i is the number of home to work trips produced in zone i. (P_i can be obtained by multiplying the zonal population, by an appropriate trip rate).

T_{ij} can be assumed to be binomially distributed, which can be approximated by a normal distribution. Thus,

$$\tilde{T}_{ij} \approx N[\, P_i \pi_{ij}, P_i \pi_{ij} (1 - \pi_{ij})\,]$$

In the following section, the optimal sample size from the standpoint of the reliability of the T_{ij} values will be investigated.

Optimal Sample Size for Trip Distribution

The optimal sample size for reliability at the distribution stage can be derived based on several alternative criteria. As discussed previously in connection with the analysis of traffic assignment values, there are two concepts of reliability that are applicable in this case—individual and simultaneous reliability. Although the sample size requirements for obtaining accuracy at the traffic assignment stage were analyzed using the concepts of both individual and simultaneous reliability, the analysis in this section will be based on the simultaneous concept only, since it is more meaningful than the other. The other alternative is related to the manner of expressing the allowable error for determining sample size requirements. The allowable error can be expressed two different ways—in terms of an absolute amount or as a percentage error. Both of these alternatives are discussed below.

Optimal Sample Size Based on Absolute Error

In the previous analysis of the reliability at the traffic assignment stage, the allowable error was expressed in terms of the absolute value 700 on the assumption that such a value has a special significance for the analysis of capacity deficiency. In the case of trip distribution, however, there is no specific absolute value of error that has any special significance, and one should select an appropriate value of δ based on the requirements of individual cases.

Using the same notations as before and assuming that the number of zones $= r$, the objective is to find the value n_i, the sample size of persons in home zone i, that will ensure that $(1-\alpha)$ per cent of the time, the estimated trips from i to j, $\tilde{T}_{ij}$, differ from the actual value T_{ij} by no more than δ, simultaneously for all zones i and j. Thus,

$$(1-\alpha) = P[\, |\tilde{T}_{ij} - T_{ij}| \leqslant \delta \text{ for all zones i and j}]$$

$$= P[\max_{1\leqslant i\leqslant r}\ \max_{1\leqslant j\leqslant r} |\tilde{T}_{ij} - T_{ij}| \leqslant \delta]$$

In order to achieve the simultaneous reliability of at least $(1-\alpha)$, it is sufficient that the reliability of T_{ij}'s of individual zones be $(1-\alpha/r)$, where r is the number of zones. Thus,

$$(1-\alpha/r) = P[\max_{1\leqslant j\leqslant r} |\tilde{T}_{ij} - T_{ij}| \leqslant \delta]$$

[This can be proved as follows:

Assuming that the immediately preceding equation is valid and using the Law of Total Probability,

$$P[\max_{1\leqslant i\leqslant r}\ \max_{1\leqslant j\leqslant r} |\tilde{T}_{ij} - T_{ij}| \leqslant \delta]$$
$$= 1 - P[\max_{1\leqslant i\leqslant r}\ \max_{1\leqslant j\leqslant r} |\tilde{T}_{ij} - T_{ij}| > \delta]$$

$$= 1 - P\left[\bigcup_{i=1}^{r}\left(\max_{1 \leqslant j \leqslant r} |\tilde{T}_{ij} - T_{ij}| > \delta\right)\right]$$

$$\geqslant 1 - \sum_{i=1}^{r} P\left[\max_{1 \leqslant j \leqslant r} |\tilde{T}_{ij} - T_{ij}| > \delta\right]$$

$$= 1 - \sum_{i=1}^{r}\left[1 - P\left(\max_{1 \leqslant j \leqslant r} |\tilde{T}_{ij} - T_{ij}| \leqslant \delta\right)\right]$$

$$= 1 - \sum_{i=1}^{r}\left[1 - (1 - \alpha/r)\right] = 1 - \sum_{i=1}^{r} \alpha/r$$

$$= 1 - \alpha\,]$$

Since $T_{ij} = P_i \pi_{ij}$ and $\tilde{T}_{ij} = P_i \tilde{\pi}_{ij}$,

$$(1 - \alpha/r) = P\left[\max_{1 \leqslant j \leqslant r} |\tilde{\pi}_{ij} - \pi_{ij}| \leqslant \delta/P_i\right]$$

Utilizing the equation (10) in page 216 of Miller (6), δ/P_i must be equal to or less than $g\,[\tilde{\pi}_{ij}(1 - \tilde{\pi}_{ij})/t_i]^{1/2}$, where g is the two sided outer α/r^2 per cent normal cut-off,[5] and t_i is the sample size of trips in zone i. Using the maximum value of

$\tilde{\pi}_{ij}(1 - \tilde{\pi}_{ij})$, which is $1/4$,[6]

$\delta/P_i = g\,[(1/4)/t_i]^{1/2}$

Therefore, $t_i = g^2 P_i^2/(4\,\delta^2)$ . . . . . (5)

and $n_i = R g^2 P_i^2/(4\,\delta^2)$ . . . . (6)

where R is a constant, or a conversion factor, reflecting the relationship of the number of persons and the trips made by them. The sample of persons, n_i, must be sufficient to provide the sample of trips t_i, so that the desired level of reliability in estimating the trips may be achieved.

Numerical Example: The equation 6 for optimal sample size was applied to the previous numerical example. To be consistent the value of δ was assumed to be 700 and the analysis was performed for 95 per cent simultaneous reliability, that is the value of α was .05. The value of g in this case was the two sided outer $(.05/7^2) = .001$ normal cut-off. Using the normal tables, $g_{.001} = 3.28$. Thus

$n_i = R\,(3.28)^2 P_i^2/4(700)^2$

In order to be able to compare the results of this approach with that of Vaughan's, P_i's were limited to include only home to work trips and n_i's were to include the number of commuters. Thus, the value of R was 1 and P_i's were equal to h_i's given in Table 2. The derived values of the sample size of commuters are shown in Table 5.

Optimal Sample Size Based on Percentage Error

As mentioned before, an alternative to expressing the allowable error in terms of an absolute amount is to express it as a proportion of related quantities. Two alternative bases can be used to derive the proportions—the trip interchange values themselves or any zonal value. In both cases, however, the basic approach for deriving the formulae for optimal sample size is similar to that used in the previous case where the error was expressed in terms of an absolute amount. Actually, the equations for the cases of percentage error, may be obtained by making appropriate substitution in equation 6. Thus, when the allowable error is expressed as a percentage (Θ) of the expected value (T_{ij}), the optimal sample size

$$n_i(ij) = \frac{R g^2 P_i^2}{4(\Theta T_{ij})^2} = \frac{R g^2 p_i^2}{4\Theta^2 (P_i \pi_{ij})^2}$$

$$\text{or } n_i(ij) = \frac{R g^2}{4\Theta^2 \pi_{ij}^2} \quad . . . . (7)$$

To ensure an accuracy of ΘT_{ij} for all T_{ij}'s

$n_i = \text{Max}_{ij}\,[\,n_i\,(ij)\,]$ . . . . (8)

Similarly, when the allowable error is expressed as a percentage ϕ of the zonal trip production (P_i), the optimal sample size,

$$n_i = \frac{R g^2 P_i^2}{4(\phi P_i)^2}$$

Samples Based on Simultaneous Reliability of Trip Interchange Values For Home to Work Trips—Error Expressed in Absolute Quantity

Sample Size & Rate for $\delta = 700$ and 95 Per Cent Reliability Assuming $P_i = h_i$

Zone	Sample Size (Commuters)	Rate in Per Cent
1	1,028	7.5
2	1,076	7.7
3	2	0.4
4	854	6.8
5	4,690	16.0
6	6,238	18.5
7	7,261	20.0
Total	21,149	15.1

TABLE 5

$$\text{or } n_i = \frac{R\,g^2}{4\,\phi^2} \quad \text{. . . . . (9)}$$

The limitation of equation 7 is that for very small values of π_{ij}, the sample size becomes too large. A common and very significant feature of both equations 7 and 9 is that they do not include any term related to zonal characteristics, such as the zonal trip production or population. Thus the sample size based on percentage error is independent of zone size and the scope of the equations cover all trips and not just the home to work trips.

Numerical Example: The use of equation 9 may be demonstrated by assuming $\phi = .10$. For 95 per cent simultaneous reliability (i.e., $\alpha = .05$), the value of g may be obtained in the same manner as that for equation (6). Thus for seven zones, g would be the two sided outer $.05/7^2 = .001$ normal cut-off. Again, assuming $P_i = h_i$ and $R = 1$, the sample size of commuters,

$n_i = (g_{.001})^2 / 4 \times (.10)^2 =$
$(3.28)^2 / .04$
or $n_i = 269$.

Since equation 9 does not contain any zonal term, the sample size for all zones will be the same, as shown in Table 6.

Samples Based on Simultaneous Reliability of Trip Interchange Values—Error Expressed as a Percentage of Zonal Trip Production

Sample Size and Rate for Error $= 0.10\ P_i$ and 95 Per Cent Reliability

Zone	Sample Size (Commuters)	Rate in Per Cent
1	269	2.0
2	269	1.9
3	269	49.8
4	269	2.2
5	269	0.9
6	269	0.8
7	269	0.7
Total	1,883	1.3

TABLE 6

ANALYSIS OF ALTERNATIVE APPROACHES & RECOMMENDED PROCEDURE

Several alternative approaches for determining the optimal O-D sample size have been presented in the previous sections. The advantages and disadvantages of the respective techniques have also been discussed. In this section the adequacy of each approach for practical application is analyzed in order to select the most appropriate technique. The recommended approach is then examined in detail.

The traffic assignment values are the ultimate result of the travel simulation procedure and, therefore, their reliability is highly desirable. However, the analysis of Vaughan's approach based on the reliability of individual link volumes revealed some of its practical limitations. The data requirement, which includes the estimation of the values of π_{ji} and $a_{ij}(kl)$, is clearly prohibitive. In addition, the sample size requirement for the simultaneous reliability of all link volumes on a realistic network is too large to be cost-effective. Moreover, the traffic assignment models are not directly related to an O-D survey and their assumptions and associated hypotheses are likely to introduce additional error. Thus, based on these considerations, the techniques based on the reliability at the traffic assignment stage are not considered practical.

The traffic distribution stage, which involves the estimation of zone to zone trip interchange values, was found to be most appropriate for the reliability analysis of O-D survey data, primarily for two reasons. The zone to zone travel data is derived directly from the O-D survey and moreover they represent the basic travel desire in an urban area. The data requirements for all of the alternative techniques based on the reliability of trip interchange values are also minimal.

The two alternative approaches for evaluating the reliability at the trip distribution stage are related to the manner of expressing the allowable error for determining sample size requirements. Although for certain purposes, it may be desirable to express an error in relative terms, for statistical analysis, it is more meaningful to express an error in absolute terms. The use of percentage error also may lead to apparently unrealistic results. An examination of the equations 7 and 9 derived on the basis of percentage error, expressed as a proportion of zone to zone trips and zonal trip productions respectively, reveals that they are independent of the zonal population. Thus in the case of equation 9, for a given level of reliability and allowable percentage of error, the sample size in absolute value is the same for all zones irrespective of the zonal population, as shown in Table 6. This is explained by the fact that as the zonal population varies, the given percentage of error actually yields varying amounts of error in absolute terms. Thus the allowable error is smaller (in absolute terms), and the level of accuracy higher for smaller zones, requiring proportionally larger sample size. Table 6 reveals that, although the sample size is the same for all zones in absolute terms, the sampling rate varies widely depending on zone size. In view of the fact that a constant percentage error may actually yield different levels of accuracy, it is recommended that the trip interchange approach based on absolute error be used for determining O-D sample size.

Recommended Procedure and Its Sensitivity

The previous discussions and derivations are oriented to commuter population and their home to work trips, primarily for the purpose of maintaining a compatibility with Vaughan's work. However, the general approach is applicable to all kinds of trips and so is the equation 6, which is repeated below:

$$n_i = Rg^2 P_i^2 / (4\delta^2)$$

where n_i = the sample size of persons for zone i;

R = a conversion factor reflecting the relationship of t_i, the sample size of trips in zone i, and n_i;

g = the two sided outer α/r^2 per cent point of the unit normal distribution, r being the number of zones;

P_i = trip production in zone i;

and δ = allowable error (number of trips) for zone to zone trips.

Equation 6 may be simplified further as the value of R and P_i can be derived in terms of the total zonal population, H_i, and the trip rate per person, q, as shown below:

R = persons per trip = 1/trips per person = 1/q

and P_i = trips per person x zonal population = qH_i

Thus $RP_i^2 = qH_i^2$ and the above equation may be replaced by the following form:

$$n_i = g^2 qH_i^2 / (4\delta^2) \quad \ldots\ldots\ldots \quad (10)$$

The procedure implied in the above equation is simple and the data requirement is also minimal. A decision has to be made regarding the level of reliability, $(1-\alpha)$, and the allowable error δ. The number zones r and the zonal population H_i will, of course, be known and the appropriate trip rate may be estimated based on previous studies in the same area or similar other areas. For instance, if the reliability of only the home to work person trips is sought a trip rate of ⅓ person trips per person may be used. Similarly, a rate of 2.5 person trips per person may be used to estimate total person trips produced in each zone. It must be noted, however, that trips rates are different urban areas and that the rates quoted above are to be used only if no prior data are available for the urban area in question. In order to be able to use the recommended procedure judiciously, one must be able to fully appreciate the relationship and sensitivity of the sample size (n_i) with each of the independent variables and a sensitivity analysis is presented below.

Sensitivity Analysis: In order to explore how the overall sample size in an urban area may vary due to varying levels of the different parameters of the equation for optimal sample size, actual computations were made for an urban area of 140,000 population,[7] and the results are presented in Table 7. In this hypothetical exercise, the areawide sample was determined by multiplying the optimal sample size for an average zone by the total number of zones in the area. Thus Areawide Sample size = r n_i = $rg^2 qH_i^2 / (4\delta^2)$

where r = number of zones,

Overall Sample Size for Varying Zonal Scheme, Level of Reliability and Allowable Error for an Urban Area of 140,000 Population— Simultaneous Reliability

No. of Zones (r) and Ave. Zonal Population (H_i)	Level of Reliability $(1-\alpha)$	Areawide Sample Size for Varying Allowable Error (δ) $\delta = 250$	$\delta = 335$	$\delta = 500$
r = 70	.95	54,208	30,188	13,552
and	.90	50,575	28,165	12,644
H_i = 2,000	.85	48,456	29,985	12,114
r = 93	.95	43,238	24,079	10,810
and	.90	40,422	22,511	10,106
H_i = 1,505	.85	38,959	21,696	9,740
r = 140	.95	30,926	17,223	7,732
and	.90	29,111	16,212	7,278
H_i = 1,000	.85	27,973	15,578	6,993

TABLE 7

$q = 2.5$ person trips per person,
$H_i = H/r$, H being the areawide population,
and the other notations are as described in the previous examples.

The relationship of the sample size with the level of reliability is straightforward and it is quite evident from the results, as expected, that for higher levels of reliability the sample size requirements are larger. Similarly, the sample size is also larger for higher levels of accuracy which is signified by smaller allowable errors. The relationship of the sample size and the zonal population, however, is subtle and must be fully understood.

The magnitude of the zonal population for a given urban area, depends primarily on the number of zones. Since the zonal population is raised to the power 2, in the equation for optimal sample size, the areawide sample size tends to be less for larger number of zones. On the other hand, the greater the number of zones, the larger is the value of $g_{\alpha/r2}$ (reflecting simultaneous reliability), which tends to increase the sample size. The combined effect of these two opposing tendencies can be determined from the results presented in Table 7, which shows that for given levels of reliability $(1 - \alpha)$ and allowable error in absolute term (δ), the overall sample size decreases as the number of zones increases. However, one must also recognize that if a given amount of allowable error (in absolute terms), δ, is expressed as a percentage of the zonal trip production, P_i, the percentage error actually increases as the number of zones is increased.

Comparison of the Recommended Procedure With FHWA Guidelines

The FHWA guidelines for determining sample size for a home-interview O-D survey in an urban area provide only an areawide sampling rate. According to these guidelines (Table 1), the overall sample size for the urban area of 140,000 population is 17,500. This sample size, however, is not explicitly related to any specific level of reliability or accuracy. The levels of reliability and accuracy associated with the FHWA recommendations, however, can be determined by comparing the FHWA sample size with those obtained by using the recommended procedure of this study. For instance, by comparing the FHWA sample size of 17,500 with the values in Table 7, it can be concluded that in the case of the urban area of 140,000 population using 140 traffic zones, the FHWA procedure assures that there is a probability of 95 per cent that none of the estimated values of zone to zone total person trips would have an error greater than 335. An error of 335 for the trip interchange values, when the average zonal trip production is only 2,500 and the number of zones is 140, does not represent a high level of accuracy. Evidently, if higher levels of reliability and accuracy are to be achieved, the sample size must be significantly larger.

It must not be overlooked that the sample sizes in Table 7 are based on the concept of simultaneous reliability which is more meaningful than the concept of individual reliability from the standpoint of analyzing the entire pattern of travel in an urban area. The concept of individual reliability, however, may be applicable for narrower objectives and the sample size requirements in that case would be significantly less. For individual reliability, the value of 'g' in the equation for optimal sample size is the two sided outer α per cent cut-off of the standardized normal distribution. The areawide sample size based on the individual reliability of the trip interchange values for the urban area of 140,000 population was computed with 140 zones and varying levels of reliability and accuracy and the results are presented in Table 8. A comparison of the FHWA sample size of 17,500 with the values in Table 8 reveals that in the case of the urban area of 140,000 population using 140 traffic zones, the FHWA procedure provides the assurance that on an individual basis there is a probability of 95 per cent that the estimated values of zone to zone total person trips woud have an error less than 220.

CONCLUDING REMARKS

The procedure for deriving an optimal sample size for O-D surveys that is presented in this paper, is a definite improvement over the existing practice, primarily because of its explicit consideration of reliability measures. However, it must be pointed out that the sample sizes based on the recommended formula in most cases are likely to be more than adequate for the desired level of reliability and the allowable error. The equation 10 on page 216 of Miller (6), which was used to obtain the formula for the sample size, is conservative and so is the use of the maximum value of ¼ for $\tilde{\pi}_{ij}(1 - \tilde{\pi}_{ij})$, which is based on the value of ½ for $\tilde{\pi}_{ij}$. In an actual case, the maximum value of $\tilde{\pi}_{ij}$ would be much less than ½ and thus the required sample size would be significantly less. The authors are pursuing this subject further and attempting to improve the proced-

Overall Sample Size for Varying Levels of Reliability and Allowable Error For an Urban Area of 140,000 Population With 140 Zones—Individual Reliability

Level of Reliability	Areawide Sample Size for Varying Allowable Error (δ)		
$(1 - \alpha)$	$\delta = 175$	$\delta = 220$	$\delta = 250$
.95	27,440	17,363	13,446
,90	19,211	12,156	9,414
.85	14,811	9,372	7,258

TABLE 8

ure by reducing the sample size requirement.

FOOTNOTES

1 It may be noted that the simulation procedure for external travel is not as standardized as that for internal travel.

2 $a_{ij}(kl)$ is treated in the analysis as a deterministic variable in contrast with a stochastic variable.

3 The analysis was involved with home to work trips, a significant portion of which occurs during the peak hour. Thus the approximate capacity of one lane ($\approx$ 700 vehicle per/hour) was considered an appropriate value of the limiting error for estimated link volumes. However, the compatibility of the unit of 'vehicles' in relation to Vaughan's assumption, $\widetilde{T}_{ij} = h_i \widetilde{\pi}_{ij}$ is questionable.

4 It may be noted that the attraction zones (j) may include the production zone i itself. Thus π_{ii} is not necessarily zero.

5 It is assumed in the analysis that $i = j = r$. If $i \neq j$ and if $i = r$ and $j = s$, g is the two sided outer α/rs per cent normal cut-off.

6 It should be noted that the use of the maximum value ¼ is a conservative approach; for further explanation, the section on Concluding Remarks may be referred to.

7 It may be noted that the urban area in the previous examples had a commuter population of 140,000 as opposed to total population as in this case.

REFERENCES

1. Federal Highway Administration, *Guidelines for Trip Generation Analysis*, U.S. Department of Transportation, June, 1967.

2. Bureau of Public Roads, *Calibrating and Testing a Gravity Model for Any Size Urban Area*, U.S. Department of Commerce, October, 1965.

3. Bureau of Public Roads. *Traffic Assignment Manual*, U.S. Department of Commerce, June, 1964.

4. Bureau of Public Roads, *Manual of Procedures for Home Interview Traffic Study*, Revised Edition (1954), Public Administration Service, Chicago, Ill.

5. Vaughan, Rodney, *Optimal Sample Sizes for Transportation Surveys*, In Transportation Science, Operations Research Society of America, May, 1972, p. 180-194.

6. Miller, Rupert G., Jr., *Simultaneous Statistical Inference*, McGraw-Hill Book Company, 1966.