Maximum-Likelihood and Bayesian Methods for the Estimation of Origin-Destination Flows

ITZHAK GEVA, EZRA HAUER, AND UZI LANDAU

The design of traffic management schemes usually requires knowledge of the pattern of trips on the system under scrutiny. This pattern is ordinarily described by an origin-destination (O-D) flow matrix. One common task of this type of matrix is the estimation of flows between the intersection approaches on a stretch of road. Estimation is based on intersection flow counts that are supplemented by a license-plate survey. In this paper a procedure is developed to obtain the most likely O-D flow estimates by using both intersection counts and results of the license-plate survey. The procedure is described in detail on the basis of a numerical example. An earlier paper reported a method of estimation that relies on intersection counts only and does not require the conduct of a sample license-plate survey. An empirical examination is conducted to test how estimation accuracy increases when the added information from the license-plate survey is used. This examination reveals that when the supplementary license-plate survey is small, the maximum-likelihood method yields unsatisfactory estimates. This deficiency is rectified by the use of a Bayesian method. The resulting solution procedure is simple, and satisfactory estimates are produced.

A variety of transportation planning and management tasks require the knowledge of the pattern of trip flows between origins and destinations. This pattern is usually described by an origin-destination (O-D) flow matrix. One common task of this type of matrix is the estimation of flows between the intersection approaches on a stretch of road. The estimation is based on a license-plate survey that is factored up to match counts of intersection flows.

In recent years attention has been given to the problem of estimating an O-D matrix by using traffic counts as the main source of information (1-4). A recent paper (5) describes a method that departs from previous work, in that travel behavior is brought into estimation by information contained in small O-D samples obtained by a survey. It is therefore not necessary to rely on speculative microstates (as in entropy models) or to assume that actual route choice is correctly captured by available models. Rather, the purpose is to find that matrix of O-D flows that is consistent with the observed traffic counts and that is most probable in view of the O-D samples observed.

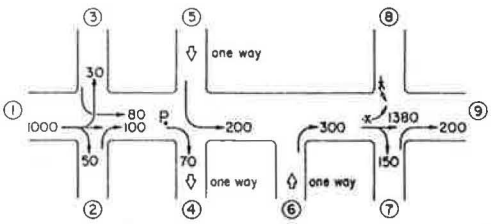
This approach is used in the present paper, in which a procedure to estimate flows between the intersection approaches on a stretch of road is developed based on intersection flow counts and a license-plate survey. The effect of sample size on estimation accuracy is explored in a real-life example.

In the first section of the paper two alternative likelihood models, which capture the manner in which data are obtained in the field, are presented. The normal equations that identify the maximum-likelihood estimate are obtained, and an algorithm for their numerical solution is described. A numerical example is presented in the second section. The example is intended to illustrate the how-to of the method and to assist the practitioner in its application. As noted earlier (6), estimates of O-D flows can be obtained from traffic counts alone, without having to resort to tedious license-plate surveys. The increase in estimation accuracy obtained as a function of sample size is examined in the third section. The results of this examination lead to the development of a new procedure based on Bayesian statistics. This procedure is presented and examined in the fourth section.

PROBLEM FORMULATION AND SOLUTION

Consider a street section as shown in Figure 1. The intersection approaches are thought of as origins and destinations. Estimates of O-D flows are desired.

Figure 1. Example of street section with eastbound flows.



The method most commonly used for this purpose in traffic engineering practice is to count traffic volumes at every intersection and to conduct a license-plate survey of a sample of vehicles entering and exiting the street of interest. Usually several digits of the license-plate number are recorded and later matched so as to obtain a sample O-D pattern. The sample is later factored up in an attempt to make the appropriate sums of O-D flow estimates match the corresponding volume counts. The purpose of this paper is to suggest an estimation procedure to replace the arbitrary and often ambiguous factoring. The merit of the procedure is that it identifies the O-D flows that are most likely in view of the results of the license-plate survey and the intersection volume counts.

In formulating the problem, the following basic notations are used:

- O_i = number of vehicles entering the street at entry approach i (i = 1,2,...,m) during a specified period of time,
- D_j = number of vehicles leaving the street at exit approach j (j = 1,2,...,n) during the same period of time,
- t_{ij} = number of license plates matched between records obtained at entry approach i and exit approach j, and
- T_{ij} = number of vehicles that enter the street by approach i and exit it by approach j.

The objective of the exercise is to obtain estimates of T_{ij} by using the data O_i, D_j, and t_{ij}. The estimation logic is of the customary maximum-likelihood kind. Thus the O-D sample matrix (t_{ij}) obtained from license-plate matching is thought to be a random sample drawn from the matrix of O-D flows (T_{ij}). The probability of observing this sample can be captured by an appropriate mathematical model. A search is made for the estimates of T_{ij} that maximize this probability and at the same time fit all the intersection volume counts. These are the most likely O-D flows to have prevailed at the time of the license-plate survey and intersection volume counts.

Two points deserve mention. First, for traffic planning and management purposes, O-D flow estimates are needed that represent average conditions rather than estimates of flows that have prevailed at the time of the survey. To do so, t_{ij}, and also O_i and D_j, would have to be regarded as random variables (7). Because the focus in this paper is the effect of the sampling ratio for the license-plate survey on O-D estimation accuracy, the estimation of O-D flows that prevailed at the time of the survey are sought. This is what practitioners have been doing anyway. The second point has to do with a discrepancy between the model and the practicalities of traffic surveys. In the model the analyst pretends that the intersection volume counts, as well as the license-plate survey, are conducted during the same time period. But because of personnel limitations, this is seldom true. With these qualifications, the random nature of the license-plate sample is described by using an appropriate probability model.

The probability model chosen must fit the manner in which the random sample is selected from the population. Thus the essential details of the license-plate survey procedure used have to be stated. To reduce survey personnel requirements and to keep errors of recording in check, it is usually best to specify beforehand some part of the license-plate number to serve as the sampling criterion. Thus if all even-numbered plates are recorded, the sampling ratio is 50 percent; if all plates ending with the digit 0 are recorded, the sampling ratio is 10 percent; and so forth. Provided that the digits selected to serve as a sampling criterion are uniformly distributed in the population of license plates, the sampling ratio is established when the sampling criterion is specified.

Two alternative probability models are suggested to capture the stochastic nature of this survey procedure. First, each license-plate match can be viewed as a success of a Bernoulli trial in which the probability of success is dictated by the sampling ratio and the rate of errors of recording and coding. The unknown flows T_{ij} correspond here to the number of Bernoulli trials. Thus the likelihood function is a product of binomial probability mass functions. Second, the license plates recorded at a certain survey point can be viewed as a random sample drawn (with replacement) from the constituent O-D flows passing that point. This leads to the multinomial probability model. Both models are considered and their merits are discussed.

Starting with the binomial model, let U_{ij} denote the number of license plates within T_{ij} that satisfy the sampling criterion. The probability distribution of U_{ij} can be described by the binomial model. Thus

$$p(U_{ij}) = \binom{T_{ij}}{U_{ij}} r^{U_{ij}} (1-r)^{T_{ij}-U_{ij}} \tag{1}$$

where r is the sampling ratio.

Equation 1 would be a reasonable description of the state of affairs if observers in the field were able to record all license plates that should be recorded and do so without error. In reality, errors occur. Thus instead of obtaining U_{ij} matching license plates for a stream of vehicles, only $t_{ij}(t_{ij} \leq U_{ij})$ is obtained.

Now the conditional probability mass function (PMF) of t_{ij} is given by

$$p(t_{ij}|U_{ij}) = \binom{U_{ij}}{t_{ij}} q^{t_{ij}} (1-q)^{U_{ij}-t_{ij}} \tag{2}$$

where q is the probability that nothing goes wrong and the license plate is obtained and processed correctly at both entry and exit.

This case is known in the literature as partial ascertainment (8). In such a case the original distribution will be distorted. If the model underlying the partial destruction of original observations (or the survival distribution) is known, the distribution of the observed values can be derived. It was shown that where the original distributions are Poisson, binomial, or negative binomial, the modified distribution is of the same form.

Therefore, the PMF of t_{ij} is also binomial and given by

$$p(t_{ij}) = \binom{T_{ij}}{t_{ij}} (rq)^{t_{ij}} (1-rq)^{T_{ij}-t_{ij}} \tag{3}$$

An expression analogous to Equation 3 can be written for every possible flow. It can be shown (9) that if $X_1,\ldots,X_t$ are binomial variates with sample size $N_1,\ldots,N_t$, respectively, and a common probability of success in each trial, then the distribution of $X = (X_1,\ldots,X_t)$ conditional on $\sum_{i=1}^{t} X_i = n$ is multivariate hypergeometric with parameters n, N, and $(N_1\ldots N_t)$. Therefore, the probability of obtaining a matrix of (t_{ij}) if the matrix of flows is (T_{ij}) is given by

$$p(t_{ij}) = \left[\prod_{i=1}^{m}\prod_{j=1}^{n}\binom{T_{ij}}{t_{ij}}\right] \Big/ \binom{\sum_i \sum_j T_{ij}}{\sum_i \sum_j t_{ij}} \tag{4}$$

The identification of the array T^*_{ij} for which this probability (or the logarithm of this probability) is maximum is needed. However, the solution must satisfy the traffic count constraints

$$\sum_{j=1}^{n} T_{ij} = O_i \quad \text{for } i = 1, 2, \ldots, m \tag{5}$$

and

$$\sum_{i=1}^{m} T_{ij} = D_j \quad \text{for } j = 1, 2, \ldots, n \tag{6}$$

By forming the Lagrangean, using Stirling's formula, taking derivatives, and equating to zero (6), the following equation is formulated:

$$T^*_{ij} = t_{ij}/(1 - A_i B_j) \quad \begin{matrix} i = 1, 2, \ldots, m \\ j = 1, 2, \ldots, n \end{matrix} \tag{7}$$

To obtain numerical values for the estimates T^*_{ij}, the unknown values $A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_n$ first must be found. This can be accomplished by a simple algorithm described in the next section.

The alternative manner of describing the survey by a probability model is to consider the random sample $t_{i1}, t_{i2}, \ldots, t_{in}$ obtained at station i as drawn from the flows $T_{i1}, T_{i2}, \ldots, T_{in}$, which are unknown. Only their sum (O_i) is given. The probability of observing this sample is given by the multinomial model:

$$\left[(\sum_j t_{ij})!/\prod_j (t_{ij}!)\right] \prod_j (T_{ij}/O_i)^{t_{ij}} \tag{8}$$

(The multinomial model is only approximate because it assumes sampling with replacement. As long as the sample is a small fraction of the population, the assumption appears proper.)

Accordingly, the probability of observing all (t_{ij}) when the matrix of flows is (T_{ij}) is given by

$$\prod_{i=1}^{m} \left\{ \left[(\sum_j t_{ij})! / \prod_j (t_{ij}!) \right] \prod_j (T_{ij}/O_i)^{t_{ij}} \right\} \tag{9}$$

The solution must satisfy the same constraints (Equations 5 and 6). By forming the Lagrangean and taking derivatives (5), the following equation is given:

$$T^*_{ij} = t_{ij}/(A_i + B_j) \qquad \begin{array}{l} i = 1, 2, \ldots, m \\ j = 1, 2, \ldots, n \end{array} \tag{10}$$

The next task is to solve a system of (m + n) simultaneous nonlinear equations with (m + n) unknowns: $A_1, \ldots, A_m$; $B_1, \ldots, B_n$. The simplest solution algorithm consists of repeated balancing of the vectors A_i and B_j and is named after Kruithof (10). The algorithm is described and illustrated by a numerical example in the following section.

NUMERICAL EXAMPLE

To illustrate the procedure, consider the road section described in Figure 1, on which the eastbound flows are obtained from ordinary intersection counts. A license-plate survey is conducted with a sampling ratio of 50 percent (r = 0.5). To achieve this sampling ratio, only vehicles with even license numbers were recorded. The number of vehicles that were matched in the survey (t_{ij}) are shown in the upper left corner of each of the 16 cells in Figure 2.

Figure 2. O-D matrix corresponding to street section in Figure 1.

From \ To	2	3	4	7	9	Σ	A_i
1	16 / 50	10 / 30	14 / 56	28 / 79	246 / 785	1000	1.0000
2			2 / 10	5 / 15	21 / 75	100	1.0494
3			1 / 4	2 / 6	21 / 70	80	1.0189
5				8 / 22	55 / 178	200	1.0047
6				10 / 28	86 / 272	300	0.9961
7					41 / 200	200	
Σ	50	30	70	150	1580	1880	
B_j			0.7513	0.6430	0.6867		

The flows T_{12}, T_{13}, and T_{79} are 50, 30, and 200, respectively, because these values can be obtained directly from the counts. Therefore, the estimation problem consists of the 13 empty cells that have to be filled with estimates so as to satisfy the 8 row and column sums. These sums are listed under the heading Σ and obtained from the intersection counts.

The solution algorithm begins by obtaining initial estimates of A_i. A starting guess may be $A_i = 1.0$. By using these tentative values for A_i, the first estimates of each B_j can be obtained. For example, for j = 4, the sum $T_{14} + T_{24} + T_{34}$ must be 70. Thus by using Equation 7,

$$[14/(1 - B_4)] + [2/(1 - B_4)] + [1/(1 - B_4)] = 70.$$

In this case $B_4 = 0.7571$. The values of B_7 and B_9 are obtained similarly by using Equation 7 to fit the given sums of columns 7 and 9. Then new estimates for A_i are calculated from the given sums of the appropriate rows and the current estimates of B_j. The new estimates of A_i are compared with the previous ones. Unless the desired closure is attained, a new round of computations is carried out. In this example, after a few iterations, the solution in Figure 2 is reached. The values of A_i and B_j are shown in the rightmost column and the lowest row, respectively. The final estimates of T_{ij} are shown in the lower right corner of each of the 13 cells. (A listing of a FORTRAN program for this procedure is available.)

The solution for the multinomial model (Equation 10) is obtained by the same algorithm. Both models produced slightly different results, which vanish after rounding to integers. Therefore, it is immaterial which model is used for the estimation.

ESTIMATION ACCURACY AND EFFECT OF SAMPLING RATIO

One of the purposes of this work has been to explore the accuracy of estimates obtainable by the method as a function of sample size. This is done empirically by comparing estimates obtained when different sampling ratios are used with the 100 percent sample. The information was provided by a detailed license-plate survey conducted on a section of a four-lane collector road with five intersections in Toronto. In the survey four digits of the license-plate code were recorded for 2 hr. The matched license-plate records were converted into O-D flows. For this investigation, these results were considered as the true matrix. It had to be pretended first that the survey was conducted with different sampling ratios by considering only license plates ending with certain digits. Flow estimates obtained by the suggested method are then compared with the true matrix.

Estimates were obtained for different sampling ratios and also for the case of zero sample [i.e., from the traffic counts only by the method described by Hauer and Shin (11)]. The error measure chosen was the average absolute error (AAE), which is defined as follows:

$$AAE = (1/N) \sum_{ij} |T^*_{ij} - T_{ij}| \tag{11}$$

where

T^*_{ij} = estimated flow from i to j,
T_{ij} = true flow from i to j, and
N = number of nonzero cells.

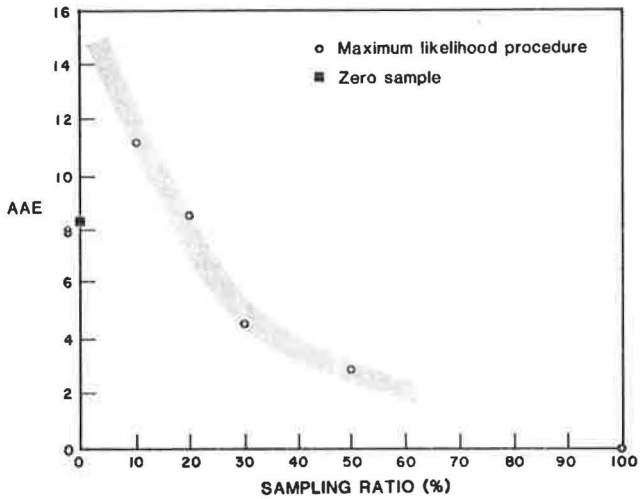
The results are shown in Figure 3 (similar results were found when other error measures were used). Some observations follow.

First, as expected, estimation accuracy increases with sample size. Initially, the improvement in accuracy is considerable. As higher sampling ratios are reached, the law of diminishing returns exerts strong influence.

Second, even without an O-D sample, reasonable flow estimates can be obtained. In this case none of the models described here can be used. The analyst has to rely on the assumption of equally likely microstates and use the method described by Hauer and Shin (6,11). The accuracy of estimation in this case (sampling ratio = 0) is shown by a square and is comparable to what can be obtained by using Equations 7 and 10 with a 20 percent O-D sample.

The reason for the unsatisfactory performance when the sample is small is inherent in Equations 7 and 10. When the flow between an O-D pair (ij) is not captured by the sample (i.e., $t_{ij} = 0$), then, of necessity, the estimate $T^*_{ij} = 0$. When the sample of license plates recorded is sufficiently small for this to occur often, estimation accuracy is likely to suffer. Thus it is not so much the sampling

Figure 3. Effect of sampling ratio on AAE of estimated matrix using the maximum-likelihood procedure.



ratio as the absolute sample size that governs estimation accuracy. When the sample size is small, the analyst can do better by ignoring it altogether because it uncovers a deficiency in the maximum-likelihood method of estimation described in the section Problem Formulation and Solution; it forces the analyst to assign zero values to flows, even though it is known that this is highly unlikely to be a satisfactory estimate. It is unwise to disregard this prior knowledge. A method that makes use of both the prior knowledge and the information contained in the O-D sample should be sought. The next section is aimed at developing such a procedure that bridges the existing discontinuity and improves estimation accuracy when relatively small samples are used.

BAYESIAN APPROACH TO ESTIMATION

The essence of Bayesian methods (12) is to apply the information contained in the outcome of an experiment to the knowledge about the probability distribution of some parameters that are available before the experiment in order to generate a new, posterior probability distribution function about these parameters.

In the present case the experiment is the license-plate survey that yields the sample realizations (t_{ij}). The prior probability distribution, denoted by $p^o(T_{ij})$, describes the probability of obtaining the matrix of flows T_{ij}. With this, and using Bayes' theorem, the posterior probability is given by

$$p(T_{ij}) \propto p(t_{ij}|T_{ij})p^o(T_{ij}) \tag{12}$$

The conditional probability component of Equation 12 has already been stated by Equation 4 (for the binomial model) or Equation 9 (in the case of the multinomial model). Thus the prior probability distribution component $p^o(T_{ij})$ must be specified.

In the absence of other information, it may be assumed that the probability of observing a certain matrix (T_{ij}) is proportional to the number of elementary events (microstates) from which it can arise (6). If all elementary events are equally likely, it can be shown that

$$p^o(T_{ij}) \propto \left[1/\left(\prod_{i=1}^{m}\prod_{j=1}^{n} T_{ij}!\right)\right] \tag{13}$$

Therefore, the posterior probability distribution function can be written as

$$p(T_{ij}) \propto \prod_{i=1}^{m}\prod_{j=1}^{n} [1/(T_{ij} - t_{ij})!] \quad \text{(binomial model)} \tag{14}$$

or

$$p(T_{ij}) \propto \prod_{i=1}^{m}\prod_{j=1}^{n} (T_{ij}^{t_{ij}}/T_{ij}!) \quad \text{(multinomial model)} \tag{15}$$

From here on, the procedure follows the logic explained in the section Problem Formulation and Solution. A search is made for that matrix T^*_{ij} that makes the posterior probability in Equations 14 and 15 as large as possible. Again, by using the method of Lagrange multipliers and Stirling's approximation,

$$T^*_{ij} = t_{ij} + A_iB_j \quad \text{(binomial model)} \tag{16}$$

and

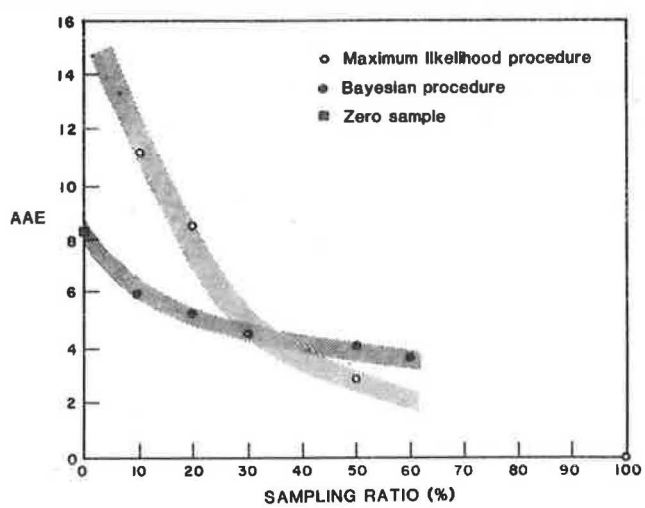
$$T^*_{ij} = \exp(t_{ij}/T^*_{ij})A_iB_j \quad \text{(multinomial model)} \tag{17}$$

Note that when $t_{ij} = 0$, both equations produce the same result ($T^*_{ij} = A_iB_j$), which is also the general solution for zero sample (6,11). In this manner the discontinuity problem near the origin (Figure 3) is eliminated.

Examination of Equations 16 and 17 reveals that the first is easily solved. Equation 17 requires a complex iterative algorithm. Both equations were used to obtain O-D estimates for the case of the street section described in the previous section. For sampling rates of up to 50 percent, both models produced almost identical estimates. For higher sampling rates, however, there is a difference between them. This can be illustrated by considering the extreme case of a 100 percent sample. At this point, Equation 16 gives the natural result $T^*_{ij} = t_{ij}$ (which is the same as Equations 7 and 10). However, Equation 17 leads to different estimates.

The effect of sampling rate on the level of accuracy, by using the maximum-likelihood procedure (Equations 7 or 10) and the Bayesian procedure (Equations 16 or 17), is presented in Figure 4. It can be seen that for sampling rates of up to 30 percent, the Bayesian method improves estimation accuracy. The maximum-likelihood procedure is appropriate for the higher sampling rates.

Figure 4. Effect of sampling ratio on AAE of estimated matrix using the maximum-likelihood and the Bayesian procedure.



SUMMARY

Two coherent methods for the estimation of O-D flows from traffic count and license-plate survey information are presented. The first estimation method identifies the most likely set of flows that agrees with the observed intersection approach flow counts on a stretch of road and the results of a sample license-plate survey.

The effect of the sample size on the accuracy of O-D flows obtained by this procedure is examined by using data from a comprehensive license-plate survey conducted on a stretch of road in Toronto. As was expected, accuracy increases with sample size. However, for small samples, better accuracy can be obtained by estimating from traffic counts only. Therefore, a second procedure based on the Bayesian approach has been developed. This procedure significantly improves the accuracy of O-D flow estimates obtained from traffic count and small sample license-plate survey information. The procedure is capable of producing relatively satisfactory estimates from small samples and thus is an aid in the performance of a common task in practice.

It appears that this procedure is preferable because of its consistency and capability, whereas the maximum-likelihood procedure should be used when high sampling rates are available at all survey stations. The Bayesian procedure described here was applied only to simple systems, such as street sections, freeway sections, and subway or bus lines. Further research is required for the application of the procedure to cases in which there are multiple paths between an O-D pair.

REFERENCES

1. L.G. Willumsen. Estimation of an O-D Matrix from Traffic Counts--A Review. Institute for Transport Studies, Univ. of Leeds, Leeds, England, Working Paper 99, 1978.
2. V.K. Chan, R.G. Dowling, and A.E. Willis. Deriving O-D Information from Routinely Collected Traffic Counts: Phase 1, Review of Current Research and Application. Institute of Transportation Studies, Univ. of California, Berkeley, 1980.
3. Y. Gur et al. Determining an O-D Trip Table Based on Observed Volumes. Presented at the Operation Research Society of America (ORSA) and The Institute of Management Science (TIMS) Annual Meeting, New York, 1978.
4. H.J. Van Zuylen and L.G. Willumsen. The Most Likely Trip Matrix Estimated from Traffic Counts. Transportation Research, Vol. 14B, 1980, pp. 281-293.
5. U. Landau, E. Hauer, and I. Geva. Estimation of Cross-Cordon Origin-Destination Flows from Cordon Studies. TRB, Transportation Research Record 891, 1982, pp. 5-10.
6. E. Hauer and B.T. Shin. Estimation of O-D Matrices from Observed Flows: Simple Systems. Univ. of Toronto/York Univ. Joint Program in Transportation, Toronto, Ontario, Canada, Res. Rept. 65, 1980.
7. H.R. Kirby and J.D. Murchland. Gravity Model Fitting with Both Origin-Destination Data and Modeled Trip Engineering Estimates. Proc., 8th International Symposium on Transportation and Traffic Theory, Toronto, Ontario, Canada, June 1981.
8. C.R. Rao. On Discrete Distribution Arising Out of Methods of Ascertainment. In Classical and Contagious Distributions (G.P. Patil, ed.), Statistical Publishing Society, Calcutta, 1965.
9. Y.M.M. Bishop, S.E. Feinberg, and P.W. Holland. Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge, Mass., 1975.
10. J. Kruithof. Calculation of Telephone Traffic. De Ingenieur, Vol. 52, No. 8, 1937.
11. E. Hauer and B.T. Shin. Origin-Destination Matrices from Traffic Counts: Application and Validation on Simple Systems. Traffic Engineering and Control, Vol. 22, No. 3, 1981.
12. L.D. Phillips. Bayesian Statistics for Social Scientists. Thomas Nelson and Sons, London, England, 1972.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.