Data Combination and Updating Methods for Travel Surveys

MOSHE BEN-AKIVA AND TAKAYUKI MORIKAWA

Data from a questionnaire survey that is the principal source of information on individual characteristics are subject to sampling errors and nonsampling biases. In aggregate data, sampling errors are usually small because a large sample is more easily obtainable and the data are also free from nonsampling biases. This paper develops a data combination and updating method that corrects survey data for nonresponse biases and reduces sampling errors by statistically combining survey data with aggregate data. This method is applied to the estimation of an origin–destination (O-D) table stratified by market segments. The O-D matrix entries and parameters of a nonresponse bias model are estimated by the maximum likelihood estimation method. An application for an intercity rail corridor is presented.

Questionnaire surveys are often used in transportation studies as a principal source of data. Most surveys obtain information from an individual in a face-to-face interview, a telephone interview, or a self-administered survey, including mail-out/mail-in, hand-out/hand-in, and hand-out/mail-in surveys. Survey data have the two following major drawbacks:

- High data collection costs and small budgets lead to small sample surveys and, consequently, large sampling errors.
- Nonsampling errors, such as nonresponse bias, result in biased statistics.

These drawbacks are often not present in data obtained by passive collection methods, such as passenger counts or data from utility records. These methods do not require the active participation of individuals and therefore convey information only on groups of individuals. Hence, this kind of data is called "aggregate data." For example, aggregate data in the form of on/off counts give only the numbers of passengers with a common origin or destination. The passengers are "grouped" with respect to origins or destinations, and the on/off counts are the column sums and row sums of the O-D table. Estimation of the cell entries must rely on additional data, such as on-board survey data, that trace each passenger's behavior.

Department of Civil Engineering and Center for Transportation Studies, Massachusetts Institute of Technology, Cambridge, Mass. 02139.

A small sample survey can be combined with aggregate data to reduce the sampling errors and correct some of the biases in the survey data. The idea is very similar to the "mixed estimation problem" in econometrics (*1*). Ben-Akiva et al. (*2*) initiated the applications of these ideas to the analysis of survey data. The combined estimation method was employed by Hsu (*3*), who estimated a transit route-level O-D table. The framework and the propositions of these methods are presented in Ben-Akiva (*4*) and are briefly reviewed in the next section.

The purpose of the present research is to develop data combination and updating methods that include the correction of survey errors in estimating O-D tables of intercity rail passengers. The survey error correction method proposed here is to combine statistically the survey data with aggregate data to obtain unbiased and more efficient estimates of unknown population parameters. The O-D tables in this application are stratified by market segments defined by trip characteristics, such as trip purpose. Data sources to be combined are on-board travel surveys and monthly ticket sales. In this application, the principal "parameters" to be estimated are cell entries of O-D matrices by market segment. Response rate parameters specific to both the market segments and the surveys are simultaneously estimated. In addition, weekend factors and monthly seasonal adjustment factors are also estimated.

REVIEW OF DATA COMBINATION AND UPDATING METHODS FOR O-D TABLE ESTIMATION

Estimation of O-D tables is one of the most important elements in transportation planning studies. Common approaches include the direct estimation of the O-D tables by using, for instance, the aggregate gravity model (*5*) or a simulation procedure based on a disaggregate destination choice model (*6*). However, these methods require large-scale surveys and complex model estimation and application procedures.

If an O-D table exists from a previous time or a preliminary estimate is available from a previous study, it can be updated or corrected using less expensive information, such as traffic counts. Here, the concept of the mixed

estimation method is used; the unknown parameters to be estimated are the cell entries of the O-D table. Survey data that provide initial estimates of the entries are referred to as "direct measurements" because they give "direct" information on the cell entries. Survey data may be collected by a passenger survey for transit or by a license plate survey for car travel.

Besides the information on every cell entry, aggregate data are also available. For example, data on the column sums and row sums of an O-D matrix are available. On/off counts at bus stops or traffic counts at the cordon line provide such data. This type of information is called "indirect measurement" because it provides "indirect" information on the cell entries.

Given these two types of information, several estimation methods are proposed in the literature. These methods attempt to maximize the "goodness" of fit (or, equivalently, to minimize the "badness" of fit) between the direct and indirect measurements. These methods also include nonstatistical approaches, such as entropy maximization (*7*) or information minimization (*8*).

The methods that have been applied most widely to the estimation of O-D tables are described next.

Iterative Proportional Fitting (IPF)

The Iterative Proportional Fitting (IPF) method is the easiest procedure to apply and has been used in many fields. In the transportation literature it has been referred to as biproportional fitting (*9*), Furness or Fratar procedure (*10*), Kruithof's algorithm (*11*), and Bregman's balancing method (*12*).

The IPF estimators are proportional to the initial matrix entries with a constant of proportionality for every row and every column. These multiplicative factors modify the initial entries to be consistent with the observed row and column sums. This method is computationally advantageous but cannot be extended appropriately to a case with stochastic indirect measurements.

Constrained Generalized Least Squares (CGLS)

The Constrained Generalized Least Squares (CGLS) estimation method is presented in Theil (*13*) and was used to estimate O-D tables by McNeil (*14*) and Hendrickson and McNeil (*15*). This method is based on the assumption that an initial matrix is equal to its corresponding true value plus a random disturbance. The disturbances usually are assumed to be normally distributed with mean zero. In other words, the initial matrix is assumed to be an unbiased estimator of the true matrix. This is a restrictive assumption because the direct measurements may have distinctive bias patterns that vary among O-D pairs.

The true matrix is estimated by the generalized least squares method subject to the equality constraints of the estimated and observed row and column sums (indirect measurements). The CGLS estimator is biased for small cell entries because of the nonnegativity constraints on the cell entries.

Constrained Maximum Likelihood Estimation (CMLE)

Although the Maximum Likelihood Estimation (MLE) method requires specific assumptions on the distribution of the observed values, it is very flexible in the sense that any type of distribution or model specification is feasible. In general, direct measurements are assumed to be independently distributed. In other words, each cell entry observation is collected by an independent sampling process.

Usually, the indirect measurements are assumed to be nonstochastic. In this case, the problem becomes the constrained maximum likelihood estimation (CMLE). Landau, Hauer, and Geva (*16*); Geva, Hauer, and Landau (*17*); and Ben-Akiva, Macke, and Hsu (*18*) applied the CMLE to O-D table estimation.

The GLS and MLE methods can also be applied with stochastic indirect measurements. If statistical independence can be assumed between direct and indirect measurements, the likelihood function to be maximized is simply the product of the likelihood of the direct and indirect measurements. The O-D table estimation method proposed in this paper is in this category, which seems most flexible. Maximum likelihood estimators are consistent and asymptotically efficient (*13*).

Geva, Hauer, and Landau (*17*) also mention that a Bayesian approach could be applied to the O-D matrix estimation if reasonable prior distributions are specified. They suggest estimators that are found by maximizing the likelihood of the posterior distribution.

DEVELOPMENT OF AN O-D TABLE ESTIMATOR WITH SURVEY ERROR CORRECTION

In the previous section the concepts of direct and indirect measurements were introduced. A maximum likelihood estimator is obtained by statistically combining both types of measurements. Direct measurements are often micro-level items and can be obtained only through survey data. Survey data, however, are subject to nonsampling errors (*19,20*). Most types of indirect measurements are obtained from other types of data collection methods and are usually free from nonresponse bias. The survey error correction method proposed here is part of an estimator in which survey data are statistically combined with aggregate data to obtain unbiased and efficient estimates of unknown population parameters.

Although the data combination methods described can be used to correct any kind of survey errors, the development in this paper is focused on the nonresponse bias that dominates all other types of errors in travel surveys. Note that this assumption simplifies the following presentation of the framework but does not imply a loss of generality.

O-D trip tables stratified by market segments are often necessary because different market segments respond differently to changes in levels of transportation services. On-board surveys can provide direct estimates of O-D tables by market segment for a transit service. However, in small samples these direct measurements have large sampling errors and may be subject to large biases due mainly to nonresponse. Ticket sales data, which are usually available for an intercity rail passenger service, provide an estimate of the aggregate O-D table that is free from nonresponse bias. The statistical combination of passenger survey data with ticket sales data can, therefore, be used to yield unbiased and more precise estimates of O-D tables stratified by market segments.

In this application the principal unknown parameters are the number of passengers belonging to a certain market segment and traveling between an O-D pair during a specific time period.

Notation

T_{ijk} = mean value of daily trips from origin i to destination j made by members of market segment k. These are the principal unknown parameters to be estimated.

p_{ijks} = response rate to survey s by individuals in market segment k traveling from i to j. These also are unknown parameters.

t_{ijks} = observed number of trips from i to j by market segment k in survey s. These are the *direct measurements.*

r_{ijm} = number of tickets sold for trips from i to j during month m. These are the *indirect measurements.*

Distributional Assumptions

Individuals in a market segment make trips according to an identical and independent Poisson process with parameter T_{ijk}. Namely, a random variable N_{ijk}, which represents the number of trips by market segment k from i to j during a randomly selected day, is the outcome of a Poisson process with parameter T_{ijk}, as follows:

$$N_{ijk} \sim \text{Poisson}\ (T_{ijk}) \tag{1}$$

Individual response to an on-board survey is the outcome of a Bernoulli trial. Individuals in a market segment have the same response rate, P_{ijks}, to survey s. Under this assumption, the direct measurement, t_{ijks}, given N_{ijk}, is a binomial random variable with parameters N_{ijk} and p_{ijks}, as follows:

$$t_{ijks} \sim \text{Binomial}\ (N_{ijk}, P_{ijks}) \tag{2}$$

The compound distribution of t_{ijks} given T_{ijk} is found by deriving the marginal distribution of t_{ijks}. From Equations 1 and 2 it can be shown that t_{ijks} has a Poisson distribution with parameter $p_{ijks}T_{ijk}$, as follows:

$$t_{ijks} \sim \text{Poisson}\ (p_{ijks}T_{ijk}) \tag{3}$$

The number of trips made in a day is statistically independent of the number made on any other day. Also, the number of trips is statistically independent among market segments.

This assumption implies that the indirect measurements, r_{ijm}, are also Poisson random variables because the sum of independent Poisson variables is Poisson distributed. Since r_{ijm} is aggregated through market segments and days of the month, it is given by

$$r_{ijm} \sim \text{Poisson}\left(d_m \sum_k T_{ijk}\right) \tag{4}$$

where d_m denotes the number of days in month m.

The survey data and the monthly ticket sales data are statistically independent. This assumption is valid if the survey data are collected on very few days in the month. The overall likelihood function is then a product of the likelihood of the survey data and that of ticket sales data.

Response Rate Specification

In the preceding equations, the response rates, p_{ijks}, are also unknown parameters. The number of these parameters can be reduced by expressing the response rate as a parametric function of the passenger's socioeconomic characteristics and the survey administration method. In this application, because of the on-board survey administration, the trip length is expected to affect the survey response rate. Assuming that a market segment is homogeneous with respect to the response rate, the following is a reasonable specification:

$$p_{ijks} = \frac{1}{1 + \exp(a_{ks} - b_{ks}d_{ij})} \tag{5}$$

where

d_{ij} = travel time from i to j; and
a_{ks}, b_{ks} = unknown parameters.

Equation 5 employs a logistic form that bounds the response rate between 0 and 1.

Under the preceding assumption, the likelihood function can now be written as a function of the data and the unknown parameters. The function is composed of two parts: direct measurements and indirect measurements. The estimated values are obtained by applying a numerical maximization algorithm.

CASE STUDY: DATA, MARKET SEGMENTATION, LIKELIHOOD FUNCTION, AND ESTIMATION TECHNIQUE

The O-D table estimation method just described is applied to the estimation of intercity rail passenger O-D tables of the Los Angeles-San Diego (LOSSAN) corridor.

Data

The Orange County Transportation Commission (OCTC) conducted on-board surveys on the following days in July 1984: 10 (Tuesday), 11 (Wednesday), 13 (Friday), and 15 (Sunday). The survey administration method and exploratory data analyses are well documented by OCTC (*21*). Amtrak also carried out an on-board survey in December of the same year. Although the questionnaire used in the survey is available, the administration method and exact date on which it was performed are unknown.

In the OCTC surveys, four out of eight trains were chosen in each direction on each day to conduct the survey. Since the combinations of those four trains differ from day to day and O-D flows dramatically fluctuate according to the combination of trains, survey data on any particular day do not necessarily mirror the daily ridership. Therefore, the OCTC three weekday surveys are combined into a single data set.

Accordingly, the following three surveys are considered:

- survey l—combined three weekday surveys by OCTC
- survey 2—survey conducted on Sunday by OCTC
- survey 3—Amtrak survey

Besides the on-board survey data, monthly ticket sales data from October 1981 through September 1985 are available. These provide the indirect measurements.

Market Segmentation

The specification of the market segmentation scheme should also depend on statistical considerations. These include a requirement of a minimal number of observations per cell and an a priori expectation with respect to response rate pattern. Namely, all members of a market segment must have approximately the same response rate and mean value of the number of trips by O-D pair. In terms of using the results for impact studies of policy changes, for example, it is desirable for a market segment to have homogeneous elasticities to the services.

The market segmentation scheme employed in this case study relies on trip purpose and the size of the traveling party as follows:

- market segment 1—commuting trips
- market segment 2—other business-related trips
- market segment 3—personal trips, traveling alone
- market segment 4—personal trips with a traveling party numbering two or more

Note: school trips are included as personal trips.

Since the Amtrak survey does not ask for the party size, it provides only an aggregate measurement of market segments 3 and 4. In other words, it provides indirect measurements of market segments 3 and 4.

Likelihood Function

In addition to the principal and the bias parameters, the model includes weekend and seasonality adjustment factors. The weekend factors are the ratio of a weekday to a weekend day value and are specific to market segment. Ridership also drastically fluctuates by season, and its fluctuation pattern depends on the O-D pair. Hence, all the O-D pairs are categorized into the following three O-D groups, and monthly seasonality factors are defined specific to each group:

- O-D group 1—O-D pairs with either the origin or the destination at Anaheim (the location of Disneyland)
- O-D group 2—O-D pairs with either the origin or the destination at San Clemente (a popular summer resort)
- O-D group 3—all the other O-D pairs

Reflecting the specification of the response rate and the weekend and seasonality factors, the model described below is obtained.

$$t_{ijks} \sim \text{Poisson}\left[\frac{1}{1+\exp(a_{ks}-b_{ks}d_{ij})}\, c_{ijm},\ w_{ks}T_{ijk}\right] \tag{6}$$

and

$$r_{ijl} \sim \text{Poisson}\left[c_{ijm}\left(E_m \sum_k T_{ijk} + F_m \sum_k w_k T_{ijk}\right)\right] \tag{7}$$

where

T_{ijk} = mean value of daily trips from i to j by market segment k (*principal unknown parameter*);
t_{ijks} = number of *respondents* of market segment k traveling from i to j in survey s (*direct measurement*);
r_{ijl} = number of tickets sold for trips from i to j in lth monthly ticket sales data (*indirect measurement*), $l = 1, \ldots, 48$ (i.e., four years);
a_{ks}, b_{ks} = unknown response rate parameters;
w_k = weekend factor of market segment k—that is, weekend/weekday ratio;
w_{ks} = wk: if survey s is conducted on weekend, 1: otherwise
c_{ijm} = seasonality factor of O-D pairs (i, j) in month m;
E_m = the number of weekdays in month m; and
F_m = the number of weekend days in month m.

Note: m_s and m_l denote the month of survey s and the lth ticket sales data, respectively.

The likelihood function to be maximized is given by

$$L = \prod_s \prod_i \prod_j \prod_k \Pr(t_{ijks}) \times \prod_l \prod_i \prod_j \Pr(r_{ijl}) \tag{8}$$

ESTIMATION RESULTS

In discussing the estimation results, it is useful to summarize first the unknown parameters. The 341 parameters

are composed of:

288 T_{ijk}'s: since LOSSAN corridor has 9 stations, one O-D table has 72 cells (9 × 8) and there are 4 market segments (4 × 72 = 288);
8 a_{ks}'s: 4 market segments times 2 surveys (OCTC and Amtrak);
8 b_{ks}'s: ditto;
4 w_k's: 4 market segments; and
33 c_{ijm}'s: 3 O-D groups times 11 months because the parameters' values for December are normalized to one.

The estimation results of the principal parameters, T_{ijk}'s, and the other parameters are shown in Tables 1 through 6. Most of the parameters have large t-statistics.

Figures 1 and 2 compare the response rates. They show that all the response rates except for market segment 1 in the OCTC survey have upward slopes. Upward slopes mean that passengers traveling longer respond to the survey more, which, intuitively, is reasonable. The reason that market segment 1 in the OCTC survey shows the downward slope seems as follows. The observed commuting trip distribution in the OCTC surveys is quite different from the distribution in the Amtrak survey. According to the OCTC surveys, the majority of the commuting trips cover a short distance—for instance, between Los Angeles and Fullerton. The Amtrak survey, however, indicates that there are more long-distance commuting trips, such as

TABLE 1 PRINCIPAL PARAMETERS FOR MARKET SEGMENT 1—COMMUTING TRIPS

	LAX	FUL	ANA	SNA	SNC	SNT	OSD	DEL	SAN
LAX	0	20.1	0.7	13.6	16.7	0.0	11.8	15.7	29.4
		(15)	(2.7)	(14)	(13)	(2.6)	(6.9)	(6.7)	(6.7)
FUL	20.9	0	0.0	0.0	3.0	0.1	2.1	4.5	1.3
	(15)		(0.3)	(0.6)	(5.5)	(1.3)	(3.8)	(5.0)	(2.3)
ANA	0.3	0.0	0	0.0	0.6	0.1	0.6	1.9	0.8
	(1.6)	(0.3)		(0.2)	(2.7)	(1.6)	(2.6)	(4.5)	(2.0)
SNA	10.9	0.0	0.0	0	0.5	0.0	1.4	0.8	0.9
	(10)	(1.2)	(0.6)		(2.1)	(0.2)	(7.3)	(2.6)	(2.2)
SNC	20.5	8.5	0.7	2.2	0	0.0	0.6	1.4	0.0
	(13)	(9.3)	(3.1)	(7.5)		(1.4)	(2.7)	(12)	(0.1)
SNT	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0
	(1.6)	(0.7)	(1.3)	(0.6)	(0.4)		(0.6)	(0.2)	(0.5)
OSD	14.1	4.4	2.9	8.0	1.1	0.0	0	0.0	0.0
	(7.6)	(7.4)	(6.0)	(11)	(4.7)	(1.4)		(2.0)	(2.2)
DEL	31.5	5.2	5.2	3.8	2.5	0.0	0.0	0	0.0
	(10)	(5.2)	(7.4)	(7.1)	(5.8)	(1.7)	(0.3)		(2.8)
SAN	70.0	2.9	2.8	1.6	0.4	0.0	0.0	0.7	0
	(11)	(2.2)	(3.7)	(3.1)	(1.6)	(0.3)	(1.9)	(5.3)	

t-statistics in parentheses

TABLE 2 PRINCIPAL PARAMETERS FOR MARKET SEGMENT 2—OTHER BUSINESS-RELATED TRIPS

	LAX	FUL	ANA	SNA	SNC	SNT	OSD	DEL	SAN
LAX	0	21.4	3.8	9.9	7.8	0.1	7.9	13.2	22.2
		(13)	(3.8)	(5.0)	(5.4)	(1.7)	(6.6)	(9.7)	(15)
FUL	26.7	0	0.0	0.0	1.2	0.0	1.2	1.7	3.6
	(11)		(0.2)	(1.3)	(2.1)	(0.6)	(2.9)	(4.2)	(4.7)
ANA	5.5	0.1	0	0.0	0.0	0.0	0.9	1.9	2.1
	(3.7)	(1.8)		(0.8)	(0.2)	(0.3)	(2.9)	(3.2)	(4.0)
SNA	9.6	0.0	0.0	0	2.0	0.0	2.2	1.1	2.7
	(5.0)	(2.2)	(0.6)		(3.4)	(0.5)	(3.9)	(2.5)	(3.2)
SNC	14.3	2.4	6.2	0.0	0	0.0	0.7	0.6	1.9
	(6.9)	(3.3)	(15)	(0.4)		(9.1)	(1.7)	(2.1)	(2.5)
SNT	0.0	0.0	0.1	0.0	0.0	0	0.0	0.0	0.0
	(1.7)	(0.6)	(2.0)	(0.5)	(5.8)		(0.7)	(1.8)	(0.5)
OSD	13.8	3.5	0.2	6.9	0.0	0.0	0	0.0	0.0
	(8.1)	(5.4)	(1.0)	(4.7)	(0.2)	(1.6)		(0.9)	(1.8)
DEL	24.7	4.8	2.6	6.4	1.9	0.2	0.9	0	0.0
	(12)	(5.7)	(5.5)	(5.3)	(2.3)	(2.9)	(3.0)		(1.9)
SAN	28.0	2.9	3.0	6.3	1.8	0.1	0.0	0.0	0
	(13)	(4.7)	(4.7)	(7.1)	(2.8)	(2.1)	(2.0)	(2.2)	

t-statistics in parentheses

TABLE 3 PRINCIPAL PARAMETERS FOR MARKET SEGMENT 3—PERSONAL TRIPS: PARTY SIZE 1

	LAX	FUL	ANA	SNA	SNC	SNT	OSD	DEL	SAN
LAX	0	82.9	14.3	52.3	51.4	2.9	61.7	68.5	131.0
		(42)	(20)	(41)	(34)	(23)	(38)	(37)	(31)
FUL	62.8	0	0.2	0.0	21.2	2.4	20.7	19.7	38.9
	(24)		(16)	(2.6)	(27)	(24)	(36)	(31)	(23)
ANA	10.1	0.2	0	0.2	4.2	0.3	8.6	8.1	14.8
	(11)	(7.0)		(14)	(19)	(5.2)	(27)	(16)	(12)
SNA	50.9	1.8	0.2	0	7.2	2.1	23.5	23.6	48.1
	(34)	(41)	(16)		(13)	(35)	(42)	(40)	(38)
SNC	44.5	20.7	0.0	12.2	0	0.0	5.2	12.5	21.0
	(36)	(24)	(0.8)	(31)		(1.8)	(19)	(42)	(31)
SNT	2.5	2.4	0.2	0.7	0.0	0	0.2	0.3	0.5
	(38)	(26)	(1.9)	(7.5)	(1.4)		(25)	(8.3)	(9.6)
OSD	64.5	20.9	7.9	32.4	7.9	0.4	0	3.6	24.3
	(33)	(34)	(15)	(30)	(24)	(29)		(48)	(57)
DEL	53.5	23.3	4.2	28.0	11.9	0.6	2.4	0	21.9
	(24)	(25)	(8.8)	(34)	(24)	(14)	(10)		(56)
SAN	124.7	59.9	20.7	58.4	19.6	1.2	21.6	21.7	0
	(23)	(33)	(29)	(43)	(33)	(20)	(54)	(55)	

t-statistics in parentheses

TABLE 4 PRINCIPAL PARAMETERS FOR MARKET SEGMENT 4—PERSONAL TRIPS: PARTY SIZE 2+

	LAX	FUL	ANA	SNA	SNC	SNT	OSD	DEL	SAN
LAX	0	4.0	4.3	3.0	4.8	0.5	9.0	13.1	62.9
		(5.3)	(7.6)	(3.9)	(4.7)	(4.7)	(7.1)	(9.5)	(17)
FUL	15.2	0	0.0	1.5	5.0	0.6	4.4	6.3	40.0
	(7.1)		(1.4)	(39)	(6.1)	(7.3)	(6.8)	(6.7)	(20)
ANA	7.6	0.0	0	0.0	0.8	0.1	0.6	0.5	14.7
	(8.5)	(1.3)		(0.4)	(4.4)	(1.3)	(2.1)	(1.8)	(13)
SNA	10.1	0.0	0.0	0	5.0	0.2	0.3	2.4	17.8
	(7.2)	(0.8)	(0.2)		(8.5)	(9.0)	(1.3)	(4.5)	(15)
SNC	5.5	1.2	1.2	0.6	0	0.0	0.8	0.6	5.1
	(5.9)	(3.1)	(4.5)	(2.8)		(0.4)	(2.9)	(3.0)	(8.3)
SNT	0.0	0.1	0.0	0.5	0.0	0	0.0	0.1	0.2
	(1.7)	(2.1)	(0.3)	(5.1)	(0.7)		(0.6)	(2.4)	(4.2)
OSD	11.8	2.0	1.3	2.2	1.8	0.0	0	0.0	0.0
	(8.2)	(3.5)	(3.1)	(3.7)	(5.1)	(2.9)		(1.4)	(2.5)
DEL	19.1	2.2	3.3	0.8	1.7	0.0	0.0	0	0.0
	(11)	(3.2)	(6.9)	(1.9)	(5.5)	(1.8)	(0.6)		(3.8)
SAN	45.4	14.0	7.1	7.9	6.6	0.1	0.3	0.0	0
	(13)	(8.8)	(9.5)	(8.5)	(11)	(2.3)	(3.6)	(2.6)	

t-statistics in parentheses

TABLE 5 RESPONSE RATE PARAMETERS

Response Rate Parameters -- $\exp(a_{ks})$

	OCTC Surveys	Amtrak Survey
m.s. 1	0.00 (1.0)	1.59 (2.8)
m.s. 2	7.71 (5.5)	7.60 (2.8)
m.s. 3	34.6 (11.1)	23.2 (4.9)
m.s. 4	0.94 (3.8)	∞ (0.6)

Response Rate Parameters -- $\exp(b_{ks})$

	OCTC Surveys	Amtrak Survey
m.s. 1	0.22 (6.4)	2.14 (5.3)
m.s. 2	1.86 (13.0)	2.71 (5.9)
m.s. 3	1.39 (55.9)	1.90 (20.3)
m.s. 4	1.09 (20.8)	0.00 (0.6)

t-statistics in parentheses

TABLE 6 WEEKEND AND SEASONAL ADJUSTMENT FACTORS

Weekend Factors -- w_k

m.s. 1	0.10 (9.5)
m.s. 2	0.23 (10.1)
m.s. 3	1.79 (25.5)
m.s. 4	1.76 (24.5)

Monthly Seasonal Adjustment Factors -- c_{ijm}

	O-D Group 1	O-D Group 2	O-D Group 3
Jan	0.91 (78)	0.94 (38)	0.99 (426)
Feb	0.98 (80)	1.13 (40)	0.95 (423)
Mar	1.48 (87)	1.34 (41)	1.15 (443)
Apr	1.55 (88)	1.89 (44)	1.22 (447)
May	1.63 (89)	4.11 (49)	1.33 (456)
Jun	1.99 (93)	7.58 (51)	1.29 (454)
Jul	1.96 (93)	7.04 (51)	1.35 (459)
Aug	2.31 (94)	6.58 (50)	1.53 (468)
Sep	1.51 (87)	2.52 (46)	1.05 (433)
Oct	0.64 (72)	1.52 (43)	0.91 (413)
Nov	0.90 (77)	0.95 (38)	1.03 (432)

t-statistics in parentheses

from Los Angeles to San Diego, than short-distance ones. This discrepancy seems to have resulted from the choice of surveyed trains and passengers' definition of "commuting." Whatever the reason may be, the MLE ascribed this discrepancy to the response rates in order to fit the data.

Figure 2 shows that there were no responses from market segment 4 in the Amtrak survey. This is because the Amtrak survey provides only the aggregate number of passengers with regard to market segments 3 and 4.

The estimates of the weekend factors look reasonable. They show that on weekends the numbers of commuters and business passengers are 10 percent to 23 percent of weekdays, respectively. Also, on a weekend day there are about 1.8 times as many personal trips as on a weekday.

Monthly seasonal factors show considerably different seasonal fluctuation patterns among O-D groups. The factor of O-D group 2 (either origin or destination is San Clemente, a summer resort) takes as much as 7.6 in June, which means that in June 7.6 times as many people as in December travel to or from San Clemente. The users of San Clemente station, however, are only a very small portion of all the patrons (in fact, only one out of eight trains stops at that station daily). O-D group 1 (either origin or destination is Anaheim, the Disneyland station) also shows higher factors in summer than does O-D group 3 (all the other O-D pairs).

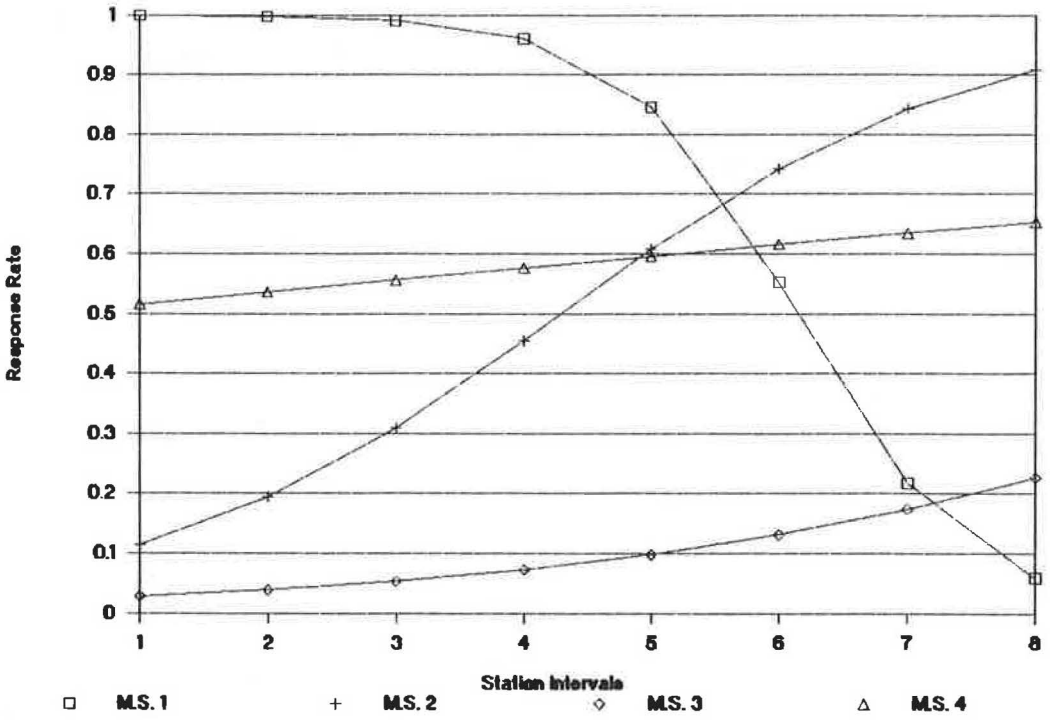

FIGURE 1 Response patterns: unrestricted model—OCTC survey.

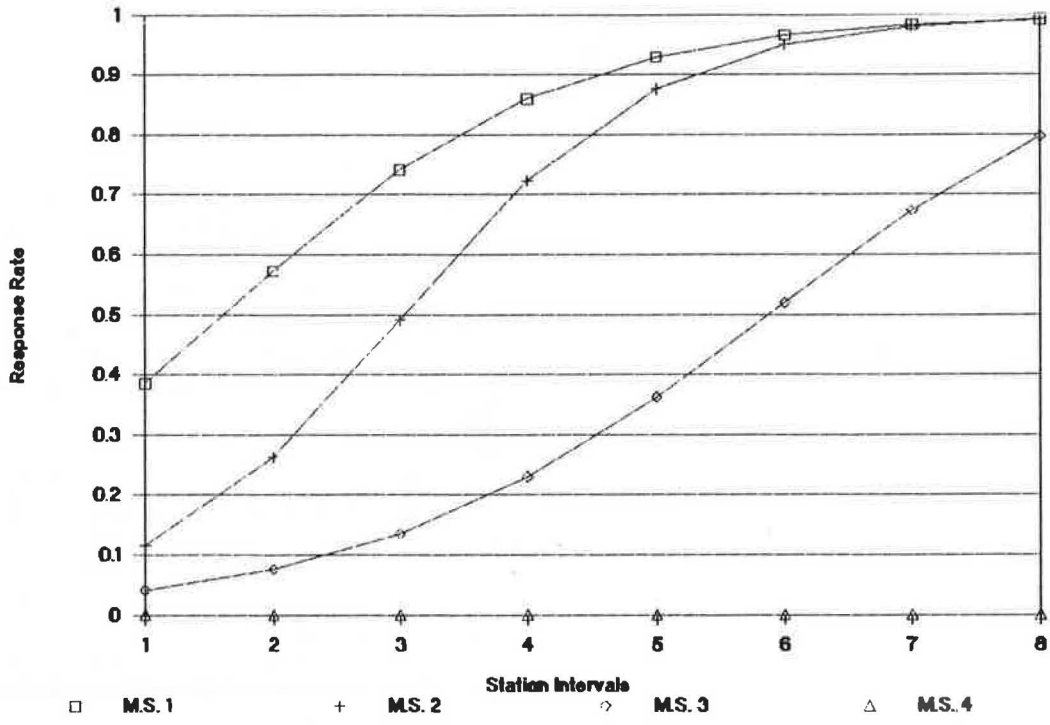

FIGURE 2 Response patterns: unrestricted model—Amtrak survey.

CONCLUSION

This paper proposes a correction method for survey errors by combining different data sources. Although reducing the survey errors by improving the survey administration or by repeating the survey may reduce the nonresponse problem, it is sometimes inefficient in terms of cost and time. Utilization of other available data sources that are inexpensive to collect seems to be an attractive and practical approach. The proposed method statistically combines survey data with aggregate data to obtain unbiased and more efficient estimates.

Although this paper focuses on the nonresponse bias, the proposed method can correct various other types of survey errors. It can also be used to update a database by treating the old database as direct measurements and combining it with up-to-date indirect measurements.

In practice, this methodology enables the utilization of any kind of information, and as much of it as possible. Conversely, it can always be used to improve future data collection methods by designing "optimal" combinations of data collection efforts.

The models described in the last section of this paper are based on some assumptions that should be investigated further. The robustness of the Poisson assumption needs to be tested by employing other distributions, such as the Normal. Secondly, alternative market segmentation schemes should be attempted. Market segmentation that properly reflects the trip generation and response behavior assumptions is required for the model.

The idea of statistically combining data from different sources can be applied in other contexts. For example, parameters of disaggregate and aggregate travel demand models can be estimated by using both survey and external counts or census data. A model transferred from one region to another is often subject to a transfer bias that can be corrected by combining data from both regions (*22*). Another application is the reweighting of survey data using the latest regional demographic data.

ACKNOWLEDGMENTS

This research was sponsored by the Federal Railroad Administration and the Transportation System Center of the United States Department of Transportation. The data for the case study were provided by Amtrak and the Orange County Transportation Commission. The authors are grateful to Stewart Butler and Arrigo Mongini, for their guidance and suggestions, and to John Prokopy, who played a central role in developing the case study. Denis Bolduc assisted with the computational aspect of this research.

REFERENCES

1. G. G. Judge, W. E. Griffiths, R. C. Hill, and T. C. Lee. *The Theory and Practice of Econometrics*. Wiley, New York, 1980.
2. M. Ben-Akiva, H. F. Gunn, and H. Pol. Expansion of Data from Mixed Random and Choice-Based Designs. Prepared for the 2nd International Conference on New Survey Methods in Transportation. Hungerford Hill, Australia, 1983.
3. P. S. Hsu. Estimation of Parameters for Multiple Spatially and Temporally Distributed Populations. Ph.D. dissertation. Massachusetts Institute of Technology, Department of Civil Engineering, Cambridge, 1985.
4. M. Ben-Akiva. Methods to Combine Different Data Sources and Estimate Origin–Destination Matrices. *Proceedings of the 10th International Symposium on Transportation and Traffic Theory*, 1987.
5. R. E. Alcaly. Aggregation and Gravity Models: Some Empirical Evidence. *Journal of Regional Science*, Summer 1967, pp. 61–73.
6. M. Ben-Akiva, H. F. Gunn, and L. Silman. Disaggregate Trip Distribution Models. *Proceedings of Japanese Society of Civil Engineers*, Vol. 347/IV-1, 1984, pp. 1–17.
7. A. G. Wilson. *Entropy in Urban and Regional Modeling*. Pion, London, 1970.
8. P. Uribe, C. G. de Leeuw, and H. Theil. *The Information Approach to the Prediction of Interregional Trade Flows*. Technical Report 6507. Economic Institute of the Netherlands, School of Econometrics, 1965.
9. M. Bacharach. *Biproportional Matrices and Input-Output Change*. Cambridge University Press, 1970.
10. S. P. Evans and H. R. Kirby. A Three-Dimensional Furness Procedure for Calibrating Gravity Models. *Transportation Research*, Vol. 8, 1974, pp. 105–122.
11. J. Kruithof. Calculation of Telephone Traffic. *De Ingenieur*, Vol. 52, 1937, pp. E15–E25.
12. B. Lamond and N. F. Siewart. Bregman's Balancing Method. *Transportation Research B*, Vol. 15B, No. 4, 1981, pp. 239–248.
13. H. Theil. *Principles of Econometrics*. Wiley, New York, 1971.
14. S. McNeil. 1983. Quadratic Matrix Entry Estimation Methods. Ph.D. dissertation. Carnegie-Mellon University, Department of Civil Engineering, Pittsburgh, 1983.
15. C. Hendrickson and S. McNeil. Estimation of Origin/Destination Matrices with Constrained Regression. *Transportation Research Record 976*, TRB, National Research Council, Washington, D.C., 1984, pp. 25–32.
16. U. Landau, E. Hauer, and I. Geva. Estimation of Cross-Cordon Origin Destination Flows from Cordon Studies. *Transportation Research Record 891*, TRB, National Research Council, Washington, D.C., 1982, pp. 5–10.
17. I. Geva, E. Hauer, and U. Landau. Maximum Likelihood and Bayesian Methods for the Estimation of O-D Flows. *Transportation Research Record 944*, TRB, National Research Council, Washington, D.C., 1983, pp. 101–105.
18. M. Ben-Akiva, P. P. Macke, and P. S. Hsu. Alternative Methods to Estimate Route Level Trip Tables and Expand On-Board Surveys. *Transportation Research Record 1037*, TRB, National Research Council, Washington, D.C., 1985, pp. 1–11.
19. W. G. Cochran. *Sampling Techniques*, 3rd ed. Wiley, New York, 1977.
20. E. D. Deming. *Some Theory of Sampling*. Wiley, New York, 1950.
21. Orange County Transportation Commission (OCTC). *An Evaluation of Amtrak's San Diegan and Metroliner Services in the Los Angeles to San Diego Corridor*. Study Report. Orange County, Calif., 1984.
22. M. Ben-Akiva and D. Bolduc. Approaches to Model Transferability: The Combined Transfer Estimation. Presented at 66th Annual Meeting of the Transportation Research Board, Washington, D.C., 1987.

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