DATA FUSION METHODS AND THEIR APPLICATIONS TO ORIGIN-DESTINATION TRIP TABLES

M. Ben-Akiva[1)], T. Morikawa[2)]

1) M. Ben-Akiva
Department of Civil Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, U.S.A.

2) T. Morikawa
Department of Transportation Engineering
Kyoto University
Sakyo-ku, Kyoto, 606, Japan

1. INTRODUCTION

Passenger and household surveys, traffic counts, revenue statistics, and population censuses provide useful data on characteristics of travelers. Each data source has its own advantages and disadvantages. Survey data and aggregate counts provide an example of two complementary data sources . Survey data offer detailed information on individual travelers characteristics, but usually have two major disadvantages: large sampling errors due to small sample sizes and nonsampling errors such as nonresponse bias. On the other hand, aggregate data from counts or other sources collected by counting or by other methods such as revenue statistics are much less costly to obtain and are usually free from sampling errors and nonsampling biases. Different types of surveys are also used to complement each other. For example, surveys of travelers crossing screen lines and cordons are often used to enrich household trip diary surveys.

Data fusion is the process of combining two or more complementary data sources into a single comprehensive data base. A data fusion method should exploit the advantages of the data sources and compensate for their disadvantages by combining them into a single data base.

The following context will elucidate the idea of combining various data sources. Data from a questionnaire survey are the principal source of information on individual characteristics. But survey data have the following two major drawbacks:

i) High data collection costs and small budgets lead to small sample sizes and consequently large sampling errors; and
ii) Nonsampling errors such as nonresponse may result in biased statistics.

These drawbacks are generally not present in data obtained by passive data collection methods such as counts of passengers. These methods do not require an active participation of individual travelers and therefore only convey information on groups of individuals. Hence, this kind of data is called "aggregate data." For example, counts of passengers boarding and alighting public transport vehicles give the numbers of passengers having a common origin or destination. The passengers are "grouped" with respect to origins or destinations. Therefore, these counts do not provide direct information on origin to destination flows and their estimation must rely on additional data sources such as an on-board survey which asks a respondent for the origin and the destination of the trip.

This paper will describe in detail an application of data fusion methods to estimate origin-destination (O-D) tables of intercity rail passengers by market segment. Ticket sales data provide monthly O-D volumes aggregated over all market segments. Estimation must, therefore, rely on on-board passenger surveys to obtain information about attributes of travelers. Such surveys are subject to significant errors. The data fusion method combines the two types of data to yield more accurate and unbiased estimates.

In this paper we present a class of methods for combining and updating data bases which correct measurement biases inherent in certain data sources and reduce sampling errors. The following section of this paper develops a general framework for combing multiple data sources and formalizes the problem as a statistical estimation process. The theoretical framework serves to lay out the range of alternative estimators. The subsequent section describes the estimation process in creating O-D tables from multiple data sources as an application of the methodology. A case study is presented in Section 4. Finally, the concluding section offers some suggestions for related areas into which the approach may extend and directions for further research.

2. GENERAL FRAMEWORK

(1) Primary unknown parameters:

Data and models are used to make inferences about characteristics of a population. In survey expansion, descriptive data analysis and forecasting, the population characteristics of interest appear as percentages of certain subgroups of a population (e.g., the percent of commuters residing in a specific zone and using public transport) or sizes of population strata (e.g., the number of trips from an origin to a destination and the number of car-owning households per specific zone). We can estimate these unknown characteristics directly, or construct and estimate an underlying parametric model predicting these characteristics in terms of explanatory variables and a few unknown parameters. In the former case, the unknown parameters are the characteristics of the population themselves. In estimating an O-D table with ten traffic zones, for example, a *non-modeling* approach must estimate 10 × 10 = 100 parameters (or cell entries), while a *modeling* approach based, for instance, on a gravity model would estimate only a small number, say 5 - 12, of model parameters. A modeling approach is usually adopted when the analysis aims to develop a model to infer the underlying behavior and apply it for forecasting. If the number of unknown parameters is prohibitively large, the use of a model in a descriptive data analysis could also be beneficial.

In this paper we focus on descriptive analyses. Hence, *the primary unknown parameters are the sizes (or percentages) of population strata (subgroups)*, which are denoted by T_k, where $k=1,...,K$ represents the stratum. [In a modeling approach we would express the primary parameters as functions using an underlying model, i.e., $T_k=f(\beta)$, $k=1,...,K$, where β is a vector of "deep" model parameters whose length is significantly less than K.]

(2) Direct measurements:

Direct measurements of the primary unknown parameters can be obtained by survey methods. For instance, the percentage of commuters residing in a given area who use a specific travel mode can be obtained by a population-based questionnaire survey. Denote by t^s_k the direct measurement of the size of the population in stratum k available from data source s. Denote by S the number of independent data sources providing direct measurements of individual strata populations. [A parametric model can be considered to be an independent data source and the values of the dependent variable obtained from a model could be viewed as direct measurements (e.g., fitted O-D flows by a gravity model).]

(3) Indirect measurements:

Indirect measurements are observations of population characteristics that depend on two or more of the primary unknown parameters. Typically, these measurements represent aggregations of population strata. For instance, if the primary parameters are cell entries of an O-D table, then traffic generation and attraction (i.e., row sums and column sums, respectively) are indirect measurements.

Indirect measurements are usually obtained from sources such as traffic counts and official census data. Denote by r_g the value of indirect measurement $g=1,...,G$, where G is the number of available independent data items providing useful information on the values of functions of two or more unknown parameters. In the vast majority of applications r_g is a linear function of the primary parameters as follows:

$$r_g = \sum_{k=1}^{K} R_{gk}T_k + v_g \ , \ g=1,...,G \ , \tag{1}$$

where R_{gk}, $g=1,...,G$ and $k=1,...,K$, are known coefficients and v_g, $g=1,...,G$, are random disturbances reflecting potential random fluctuations and measurement errors. If, for example, T_k is the number of trips from an origin to a destination and r_g is a traffic count at a given location, then the coefficient R_{gk} is the share of the volume T_k crossing counting station g.

(4) Measurement bias:

Some measurement techniques — surveys in particular — are often subject to biases originating from a variety of systematic non-sampling errors which depend among other things on survey administration procedures. The following presentation assumes, for simplicity, that the indirect measurements are unbiased, i.e., $E(v)=0$.

Let $\mu^s{}_k$ be a multiplicative systematic bias factor for the s direct measurement of the population of stratum k. Thus, the expected value of direct measurement s of stratum k's population is expressed by:

$$E(t^s{}_k) = \frac{\mu^s{}_k T_k}{F^s{}_k} \ , \ k=1,...,K, \ s=1,...,S \ , \tag{2}$$

where $F^s{}_k$ is the design expansion factor of measurement s for stratum k. With prespecified sampling rates, $F^s{}_k$ is equal to the inverse of the sampling rate of survey s in stratum k. The design expansion factors (the F's) are assumed to be known. [If for some reason they are unknown, then they would be omitted and the corresponding bias parameters would represent the inverse of "bias corrected" expansion factors.] Without loss of generality, the value of the direct measurements can be expressed as follows:

$$t^s{}_k = \frac{\mu^s{}_k T_k}{F^s{}_k} + u^s{}_k \ , \ k=1,...,K, \ s=1,...,S \ , \tag{3}$$

where $u^s{}_k$ is the random error in the measurement of $t^s{}_k$. Since $\mu^s{}_k$ is unknown, the introduction of a multiplicative bias results in a non-linearity in the unknown parameters. It appears in the above equation in the form of the product $\mu^s{}_k T_k$.

(5) Estimation method:

The problem of combining different data sources may now be stated as the task of jointly estimating the unknown parameters of equations (1) and (3). The unknown parameters are the K primary parameters denoted by the vector T and the SK bias parameters denoted by μ. The observable variables are SK observations of t's and G observations of r's, the F's and the R's are known constants and the u's and the v's are random errors. Thus, the number of unknown parameters is $K+SK$ and the number of independent measurements is $SK+G$. The necessary condition for all the unknown parameters to be identifiable is $G \geq K$. This condition is infeasible in most applications. Consider, for instance, the problem of estimating an O-D table with K cells. This condition implies that at least K independent traffic counts are needed in addition to K direct measurements of the K cells.

Thus, this estimation problem is impractical unless the number of bias parameters is reduced from SK to a much smaller number. This requires *a priori* specifications of parametric bias models which express the values of the SK μ's as functions of strata attributes and a smaller number of structural or "deep" bias parameters that need to be estimated from the data.

The joint estimation of a system of equations such as (1) and (3) is known in the econometrics literature as a *mixed estimation problem* (Judge et al (1)). If these equations were linear in the unknown parameters, they could be estimated by ordinary (or generalized) least squares methods. The introduction of unknown multiplicative bias parameters brings about non-linearities requiring the use of non-linear estimators that significantly increase the computational burden. A *Maximum Likelihood Estimator* (MLE) can be employed if the parametric form of the distribution of the t's and the r's can be specified. MLE provides full flexibility in model specification and has desirable statistical properties under very general conditions.

The generalized form of the estimation problem can be expressed as follows:

$$\underset{\{T,\mu\}}{\text{minimize}}\ h(u,v{:}T,\mu) = h^1(u{:}T,\mu) + h^2(v{:}T)\ , \tag{4}$$

where T is a $K \times 1$ vector of the primary unknown parameters and μ is the vector of the bias parameter. $h(u,v{:}T,\mu)$ is an expression of total error or "badness" of fit to be minimized. It is reasonable to assume that the direct and indirect measurements are independently distributed and the objective function is therefore additively separable into the two error functions $h^1(u{:}T,\mu)$ and $h^2(v{:}T)$.

A special case of practical significance is the situation with deterministic (i.e., error free) indirect measurements. For v=0 the estimation problem becomes:

$$\underset{\{T,\mu\}}{\text{minimize}}\ h^1(u{:}T,\mu)$$

$$\text{subject to}\quad \mathbf{r} = \mathbf{RT}\ , \tag{5}$$

where $\mathbf{r}$ is a $G \times 1$ vector of indirect measurements and R is a $G \times K$ matrix of known coefficients. If $G \geq K$, the values of the Ts can be obtained from the constraints by solving K linear equations. Thus, the estimation problem in (5) is relevant for situations in which $K > G$.

The computational advantage of this deterministic indirect measurement can be seen in the Iterative Proportional Fitting (IPF) method, which is also known as biproportional fitting (2), Furness or Fratar procedure (3), the Kruithof's algorithm (4), and Bregman's balancing method (5). 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. Deterministic indirect measurements are also employed in Constrained Generalized Least Squares (Hendrickson and McNeil, (6)) and Constrained Maximum

Likelihood Estimation (Ben-Akiva, Macke and Hsu, (7)). Alternative estimators for stochastic and deterministic indirect measurements are presented in Ben-Akiva (8) and McNeil and Hendrickson (9).

3. O-D TABLE ESTIMATION WITH SURVEY ERROR CORRECTION

This section applies the general framework to a special case. The primary unknown parameters are O-D volumes of passengers and the data sources are questionnaire surveys and counts. A unique aspect of this application is the stratification of the O-D table by market segments. The segmentation is observable in a survey but not in the count data. Note also that there are multiple counts over time which are treated as random variables.

Trip tables stratified by market segment are often necessary because different market segments respond differently to changes in the attributes of transportation services. For public transportation services on-board and platform passenger surveys can provide direct estimates of market segment trip tables. However, these direct measurements for specific O-D flows have large sampling errors and are often biased mainly due to nonresponse problems. Passenger counts and ticket sales data may provide estimates of aggregate O-D flows which are free from nonresponse bias. The combination of passenger survey data with passenger counts and ticket sales data can therefore be used to yield unbiased and more precise estimates of trip tables stratified by market segment.

In this application, the primary unknown parameters are the number of passengers belonging to a certain market segment and traveling between a certain O-D pair during a specific time period. This problem can also be defined as the estimation of the entries of a three-dimensional array in which the first axis represents trips origin, the second spans the trip destination, and the third dimension is the market segment. The passenger survey data provide a direct measurement for each entry and counts or ticket sales data give indirect measurements of the entries aggregated through the third axis. We assume, then, that counts or ticket sales data are routinely collected over time and therefore are available for survey periods as well as for other times. The direct measurements are assumed to be biased. This data combination problem also includes the estimation of the unknown parameters of an explicit response bias (or response rate) specification.

(1) Notation:

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

p_{ijks}= response rate to survey s by individuals in market segment k traveling from i to j. These are also 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 or a count of trips from i to j during month m. These are indirect measurements.

(2) Distributional assumptions:

i) Individuals in a market segment make trips according to an identical and independent Poisson process. 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:

$$N_{ijk} \sim \text{Poisson}(T_{ijk}) \qquad (6)$$

or

$$\Pr(N_{ijk}) = \frac{T_{ijk}{}^{N_{ijk}} \exp(-T_{ijk})}{N_{ijk}!} . \tag{7}$$

ii) Individual response to a survey is the outcome of a Bernoulli trial. Also 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:

$$t_{ijks} \sim \text{Binomial}(N_{ijk}, p_{ijks}) \tag{8}$$

or

$$\Pr(t_{ijks} \mid N_{ijk}) = \binom{N_{ijk}}{t_{ijks}} (p_{ijks})^{t_{ijks}} (1-p_{ijks})^{N_{ijk}-t_{ijks}} . \tag{9}$$

The compound distribution of t_{ijks} given T_{ijk} is found by deriving the marginal distribution of t_{ijks}. From equations (7) and (9) it can be shown that t_{ijks} has a Poisson distribution with parameter $p_{ijks}T_{ijk}$, that is:

$$t_{ijks} \sim \text{Poisson}(p_{ijks}T_{ijk}) \tag{10}$$

iii) The number of trips made in a day is statistically independent of any other day and 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:

$$r_{ijm} \sim \text{Poisson}(d_m \sum_k T_{ijk}) , \tag{11}$$

where d_m denotes the number of days in month m.

iv) The survey data and the monthly count data are statistically independent. This assumption is valid if there are very few survey days during any given month. The overall likelihood function is then a product of the likelihood of the survey data and that of ticket sales data.

(3) Response rate model:

In the above equations the response rates, p_{ijks}, are also unknown parameters. The number of these parameters can be reduced by expressing them as functions of the passenger's socioeconomic characteristics and the survey administration method. For a self-administered on-board survey the travel time may affect the survey response rate. Assuming that a market segment is homogeneous with respect to survey response rates, the following specification of response rate as a function of travel time may be used:

$$p_{ijks} = \frac{1}{1+\exp(a_{ks} - b_{ks}d_{ij})} , \tag{12}$$

where

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

Equation (12) employs a logistic form to bound the response rate between 0 and 1.

(4) Likelihood function:

The likelihood function of the direct measurements is:

$$L_1 = \prod_i \prod_j \prod_k \prod_s h^1(t_{ijks}\colon T_{ijk},\, a_{ks},\, b_{ks})\ , \tag{13}$$

and that of the indirect measurements is:

$$L_2 = \prod_i \prod_j \prod_m h^2(r_{ijm}\colon T_{ijk})\ . \tag{14}$$

The forms of h^1 and h^2 are based on the Poisson distributions in equations (10) and (11), respectively.

Under distributional assumption (iv), the overall likelihood function to be maximized is given by:

$$L = L_1 \times L_2\ . \tag{15}$$

4. AN EMPIRICAL CASE STUDY

This section presents an application of the O-D table estimation method developed in the previous section to the estimation of intercity rail passenger trip tables for the Los Angeles-San Diego (LOSSAN) corridor.

4.1 Data, Market Segmentation, Likelihood Function and Estimation Technique

(1) Data:

Orange County Transportation Commission (OCTC) conducted on-board surveys on the following days in July, 1984: 10 (Tue), 11 (Wed), 13 (Fri), and 15 (Sun). The survey administration method and exploratory data analyses are documented in OCTC (10). 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 the exact date have not been documented.

In the OCTC surveys, four out of eight trains were chosen in each direction on each day to administer 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, we consider the following three surveys:
survey 1 - OCTC weekday surveys;
survey 2 - OCTC Sunday survey; and
survey 3 - Amtrak survey.

In addition to the on-board survey data, indirect measurements were provided by monthly ticket sales data from October, 1981 through September, 1985.

(2) Market segmentation:

The specification of the market segmentation scheme should depend on both substantive and statistical considerations. It is clearly desirable to define market segment that are homogeneous with respect to demand elasticities with respect to service attributes. The statistical considerations include a requirement of a minimal number of observations per cell and a priori expectation with respect to response rate and travel pattern. Namely, all members of a market segment must have approximately the same response rate and mean value of trips by O-D pair.

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; and
Market Segment 4 - Personal trips with a traveling party size of two or more.
(Note: school trips are included as personal trips.)

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

(3) 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 groups and monthly seasonality factors are specific to each group:

O-D group 1 - O-D pairs with either origin or destination at Anaheim (the location of Disneyland);
O-D group 2 - O-D pairs with either origin or destination at San Clemente (a popular summer resort); and
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:

$$t_{ijks} \sim \text{Poisson}\left[\frac{1}{1+\exp(a_{ks} - b_{ks}d_{ij})}\, c_{ijm_l} w_{ks} T_{ijk}\right], \tag{16}$$

and

$$r_{ijl} \sim \text{Poisson}\left[c_{ijm_l}\left(E_{m_l}\sum_k T_{ijk} + F_{m_l}\sum_k w_k T_{ijk}\right)\right], \tag{17}$$

where

t_{ijks} = the number of respondents in survey s belonging to market segment k and traveling from i to j (*direct measurement*);

r_{ijl} = the number of trips from i to j during the l-th month, l=1,...,48 (i.e., 4 years) (*indirect measurement*);

T_{ijk} = the mean value of the number of weekday trips from i to j by market segment k (*primary unknown parameter*);

a_{ks}, b_{ks} = unknown response rate parameters for market segment k and survey s;

d_{ij} = travel time from i to j, approximated by the number of stations between i to j;
w_k = weekend factor for market segment k, i.e., weekend/weekday ratio;
w_{ks} = w_k, if survey s is conducted in weekend; 0, otherwise;
c_{ijm} = seasonality factor for O-D pairs (i,j) and 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 l-th count data, respectively.)

The log-likelihood function to be maximized is given by:

$$L = \prod_s \prod_i \prod_j \prod_k \Pr(t_{ijks}) \times \prod_l \prod_i \prod_j \Pr(r_{ijl}) . \tag{18}$$

A quasi-Newton numerical method called "Davidon-Fletcher-Powell (DFP)" (Fletcher and Powell (11)) procedure was employed to solve this optimization problem due to the large number of unknown parameters since this procedure does not require the inverse operations of the Hessian matrix. It was implemented on a personal computer using the GAUSS programming language (Aptech Systems (12)). The program required approximately 500 iterations to reach convergence.

4.2 Estimation Results

In discussing the estimation results, it is useful to summarize first the unknown parameters. The 341 unknown 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 increases with distance traveled. This means that passengers traveling longer are more likely to respond to an on-board survey, which is intuitively reasonable. Market segment 1 in the OCTC survey has a negative slope, which can be explained by the following: in the OCTC survey data, most trips are short distance (e.g., Los Angeles - Fulerton), while the Amtrak survey indicates that there are more long-distance commuting trips, such as Los Angeles - San Diego, than short distance ones. This discrepancy seems to have resulted from the choice of surveyed trains and passengers' subjective definitions of the term "commuting" that they have used in answering the survey question on trip purpose. Thus, the response rate parameter for the OCTC survey may have captured the effect of selectivity of the trips resulted from the above two reasons.

Because the Amtrak survey provides only the aggregate number of passengers with regard to market segments 3 and 4, Figure 2 shows that there were no responses from market segment 4 in the Amtrak survey due to an estimation problem.

Table 1 Primary Parameters for Market Segment 1 - Commuting Trips; persons/day (t-statistics in parentheses)

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

Table 2 Primary Parameters for Market Segment 2 -.Other Business related Trips; persons/day (t-statistics in parentheses)

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

Table 3 Primary Parameters for Market Segment 3 -.Personal Trips: Party Size 1; persons/day (t-statistics in parentheses)

	LAX	FUL	ANA	SNA	SNC	SNT	OSD	DEL	SAN
LAX		82.9 (42)	14.3 (20)	52.3 (41)	51.4 (34)	2.9 (23)	61.7 (38)	68.5 (37)	131.0 (31)
FUL	62.8 (24)		0.2 (16)	0.0 (2.6)	21.2 (27)	2.4 (24)	20.7 (36)	19.7 (31)	38.9 (23)
ANA	10.1 (11)	0.2 (7.0)		0.2 (14)	4.2 (19)	0.3 (5.2)	8.6 (27)	8.1 (16)	14.8 (12)
SNA	50.9 (34)	1.8 (41)	0.2 (16)		7.2 (13)	2.1 (35)	23.5 (42)	23.6 (40)	48.1 (38)
SNC	44.5 (36)	20.7 (24)	0.0 (0.8)	12.2 (31)		0.0 (1.8)	5.2 (19)	12.5 (42)	21.0 (31)
SNT	2.5 (38)	2.4 (26)	0.2 (4.9)	0.7 (7.5)	0.0 (1.4)		0.2 (25)	0.3 (8.3)	0.5 (9.6)
OSD	64.5 (33)	20.9 (34)	7.9 (15)	32.4 (30)	7.9 (24)	0.4 (29)		3.6 (48)	24.3 (57)
DEL	53.5 (24)	23.3 (25)	4.2 (8.8)	28.0 (34)	11.9 (24)	0.6 (14)	2.4 (10)		21.9 (56)
SAN	124.7 (23)	59.9 (33)	20.7 (29)	58.4 (43)	19.6 (33)	1.2 (20)	21.6 (54)	21.7 (55)	

Table 4 Primary Parameters for Market Segment 4 -.Personal Trips: Party Size 2+; persons/day (t-statistics in parentheses)

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

Table 5 Response Rate Parameters
(t-statistics in parentheses)

Response rate parameters - exp(a_{ks})

Market segment	OCTC survey		Amtrak survey	
1	0.00	(1.0)	1.59	(2.8)
2	7.71	(5.5)	7.60	(2.8)
3	34.6	(11.1)	23.2	(4.9)
4	0.94	(3.8)	∞	(0.6)

Response rate parameters - exp(b_{ks})

Market segment	OCTC survey		Amtrak survey	
1	0.22	(6.4)	2.14	(5.3)
2	1.86	(13.0)	2.71	(5.9)
3	1.39	(55.9)	1.90	(20.3)
4	1.09	(20.8)	0.00	(0.6)

Table 6 Weekend and Seasonal Adjustment Factors
(t-statistics in parentheses)

Weekend Factors - w_k

Market segment		
1	0.10	(9.5)
2	0.23	(10.1)
3	1.79	(25.5)
4	1.76	(24.5)

Monthly seasonal adjustment factors - c_{ijm}

	O-D group 1		O-D group 2		O-D group 2	
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)

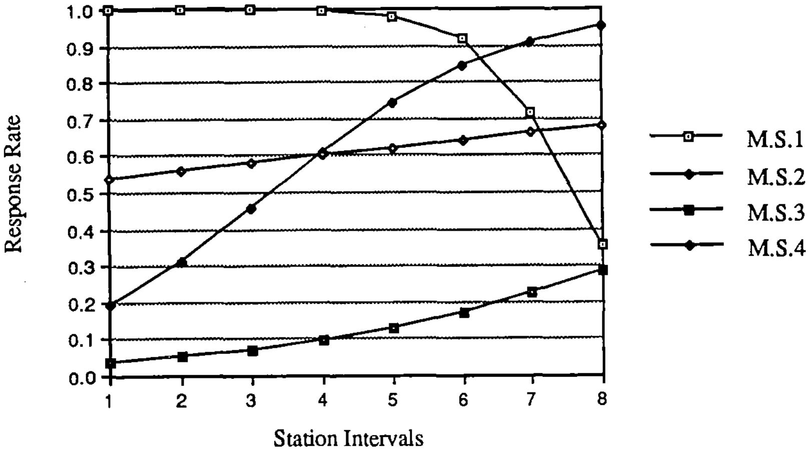

Figure 1 Response Patterns in OCTC Surveys

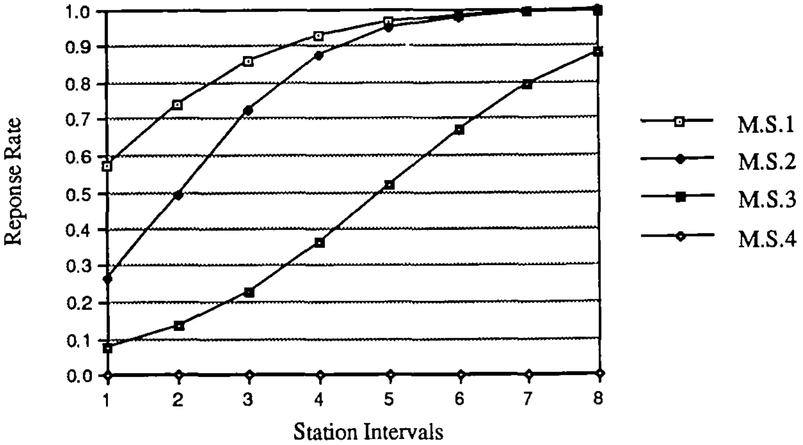

Figure 2 Response Patterns in Amtrak Survey

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

Monthly seasonal factors show considerable variation among O-D groups. The factor of O-D group 2 (either origin or destination is San Clemente, a summer resort) is 7.6 in June, which means that in June 7.6 times as many travelers as in December travel to or from San Clemente. Note, however, that the journeys to and from San Clemente station are a trivial fraction of the total ridership (in fact, only one out of eight trains stops at that station). O-D group 1 (either origin or destination is Anaheim, the Disneyland station) also shows greater factors in summer months than O-D group 3 (all the other O-D pairs).

5. CONCLUSIONS

The method described in this paper can be used in a variety of applications. The application presented focused on the correction of the nonresponse bias and reduction of sampling errors by combining survey data and aggregate count data. Although improving the survey administration or repeating the survey may reduce survey errors and biases, it may not be efficient in terms of cost and time. Utilizing other sources of information provided by existing data or inexpensive counts seems feasible and practical. The proposed method statistically combines survey data with aggregate data and obtains unbiased and efficient parameters estimates.

The idea of statistically combining data from different sources can be applied to various other contexts. In this paper we focused on the use of data to obtain descriptive statistics of an existing situations. Another area where the same ideas seem to have a significant potential is in the estimation of travel demand models. Parameters of disaggregate and aggregate travel demand models can be estimated by using both survey and external counts or census data (e.g., Gonzalez (13), Morichi and Yai (14)). 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 (Ben-Akiva and Bolduc (15)). Combining revealed and stated preference data in the estimation of travel demand model is another promising area of research and can fit within the general framework presented in this paper (Morikawa (16)).

ACKNOWLEDGMENTS

This research was partially supported 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. John Prokopy from Amtrak was instrumental in obtaining the data and formulating the case study.

REFERENCES

1. Judge, G.G., Griffiths W.E., Hill, R.C. and Lee, T.C. (1980), The Theory and Practice of Econometrics, Wiley, New York.

2. Bacharach, M. (1970), Biproportional Matrices and Input-Output Change, Cambridge University Press.

3. Evans, S.P. and Kirby, H.R. (1974), A Three-Dimensional Furness Procedure for Calibrating Gravity Models, Transportation Research, 8, pp. 105-122.

4. Kruithof, J. (1937), Calculation of Telephone Traffic, De Ingenieur, 52, E15-E25.

5. Lamond, B. and Stewart, N.F. (1981), Bregman's Balancing Method, Transportation Research B, Vol. 15B, No. 4, pp. 239-248.

6 Hendrickson, C. and McNeil, S. (1984), Estimation of Origin/Destination Matrices with Constrained Regression, Transportation Research Record, 976, pp. 25-32.

7. Ben-Akiva, M., Macke, P.P. and Hsu, P.S. (1985), Alternative Methods to Estimate Route Level Trip Tables and Expand On-Board Surveys, Transportation Research Record, 1037, pp. 1-11.

8. Ben-Akiva, M. (1987), Methods to Combine Different Data Sources and Estimate Origin-Destination Matrices, in Transportation and Traffic Theory (Proceedings of the 10th International Symposium on Transportation and Traffic Theory), N.H. Gartner and N.H.M. Wilson eds., Elsevier, pp. 459-481.

9. McNeil, S. and Hendrickson, C. (1985), A Note on Alternative Matrix Entry Estimation Techniques, Transportation Research, 19B, pp. 509-519.

10. Orange County Transportation Commission (OCTC) (1984), An Evaluation of Amtrak's San Diegan and Metroliner Services in the Los Angeles to San Diego Corridor, Study Report, Orange County, California.

11. Flethcer, R. and Powell, M.J.D. (1963), A Rapidly Convergent Descent Method for Minimization, Computer Journal, pp. 401-408.

12. Aptech Systems, Inc. (1988), GAUSS - Version 2.0 Manual, Kent, Washington, U.S.A.

13. Gonzalez, S.L. (1985), Combining Survey and Aggregate Data for Model Estimation, Ph.D. Dissertation, Department of Civil Engineering, MIT.

14. Morichi, S. and Yai, T. (1988), Estimation of Disaggregate Model Using Additional Aggregate Data, Proceedings of PTRC Annual Meeting.

15. Ben-Akiva, M. and Bolduc, D. (1987), Approaches to Model Transferability and Updating: The Combined Transfer Estimation, Transportation Research Record, 1139, pp. 1-7.

16. Morikawa, T. (1989), Incorporating Stated Preference Data in Travel Demand Analysis, Ph.D. Dissertation, Department of Civil Engineering, MIT.