CHAPTER 3

FORECASTING THE VALUES OF EXOGENOUS VARIABLES: SOCIOECONOMIC VARIABLES

Introduction

Disaggregate behavioral model forecasts of the effects of urban transportation policy require auxiliary forecasts of the variables exogenous to the model system. The exogenous variables typically include residence and work location, and household socioeconomic and demographic characteristics. Consistent aggregation of behavioral models requires that these variables be provided for each homogeneous market segment, or for a representative random sample of households. The forecasting method should take into account shifts in demographic and land use patterns, changing economic conditions, and population growth.

It is in the nature of auxiliary forecasting that one does not have available complete structural or causal models; hence, forecasting must use data analysis and trend projection techniques, combined with available external forecasts. The method should be able to combine the information contained in a variety of different data sources, and have the capacity to upgrade the quality of the forecasts as additional data become available.

We have developed a methodology--SYNSAM--for generating a synthetic representative sample of households for an urban area for any specified date.[1] We describe the implementation of this procedure for the San Francisco Bay Area, involving the construction of a sample of 12,000 households for the year 1976. In

[1]This methodology is described in detail, with technical specifications and a description of the software, in Final Report volume VIII.

addition to residence and work locations, data for each household comprises a subset of the socioeconomic variables tabulated in the Public Use Sample (PUS) of the 1970 Census. The implementation utilizes 1960 and 1970 Census data plus external projections of population and economic conditions. Because such data are available for all Standard Metropolitan Statistical Areas (SMSA), the procedure is readily transferable to other cities.

A principal feature of the SYNSAM procedure is the use of Iterative Proportional Fitting (IPF) to construct and update for each zone of residence a contingency table giving the distribution of a selected set of household characteristics, starting from the various marginal tabulations available on census tapes and other sources. The program is based on an algorithm due to Haberman (1974). A second principal step is to actually construct a synthetic sample by random sampling, once the contingency tables for socioeconomic characteristics have been computed. For each household in the sample, the program selects a residence zone; selects a vector of nine household socioeconomic characteristics; assigns an employment zone; selects a matching representative household with the same vector of socioeconomic characteristics from the PUS census file; and selects a worker within this household.

SYNSAM is intended to be a flexible methodology, capable of accommodating a variety of data sources and estimates, and allowing alternative methods and possible improvements in various steps of the procedure.

Background: Construction of Sampling Tables

Exogenous socioeconomic variables are normally defined or coded categorically; we shall assume they always have this form. Then the distribution of characteristics-of-household is described by a contingency table, such as the schematic table given in Figure 12. Each cell in the table represents a "homogeneous market segment," and the cell probability gives the share of this market segment in the population. The dimensions of the table correspond to the exogenous variables, including residence and workplace zone. For each zone pair, the transportation networks for the urban area provide level of service (LOS) variables, including times and costs, for alternative transportation modes. The disaggregate travel demand model system forecasts the travel behavior of the households in a cell as a function of the cell socioeconomic characteristics and LOS variables. In a complete model, this will include trip generation and distribution and mode split probabilities. The sum of these probabilities over cells, weighted by all probabilities, gives aggregate travel behavior forecasts for the urban area. In the special case that the only exogenous variables are residence and workplace zone, the system above reduces to the conventional aggregate demand forecasting framework; hence, the auxiliary forecasting methodology considered here is applicable to aggregate as well as disaggregate policy analysis.

In Figure 12, cell probabilities are denoted P_{ij}, where i is the level of the first variable, and j is the level of the second.

FIGURE 12 Example of a Contingency Table

Number of Persons in Household	Residence Zone 1	2	3	SUM
1	p_{11}	p_{12}	p_{13}	p_{1+}
2	p_{21}	p_{22}	p_{23}	p_{2+}
3	p_{31}	p_{32}	p_{33}	p_{3+}
4	p_{41}	p_{42}	p_{43}	p_{4+}
5	p_{51}	p_{52}	p_{53}	p_{5+}
SUM	p_{+1}	p_{+2}	p_{+3}	$p_{++}=1$

The marginal probabilities are denoted $p_{i+} = p_{i1} + p_{i2} + p_{i3}$ for the number of persons in household and $p_{+j} = p_{1j} + p_{2j} + p_{3j} + p_{4j} + p_{5j}$ for residence zone. In practice, tables will usually be required for more than two variables; in the present application, nine-way contingency tables are constructed for the nine household variables given in Table 12.

Cell probabilities are denoted by P_σ, where the index is a vector $\sigma = (i_1,...,i_9)$ of these nine socioeconomic and demographic variables. Because of changes over time in population and in demographic and socioeconomic characteristics, the cell probabilities will be functions of time. When the dates of probabilities must be distinguished, we write $p_\sigma(t)$ for date t.

The basic problem is that data for the full contingency table is rarely, if ever, available. It could be obtained directly only from a large-scale random survey of households in the area. Without such a survey, the available data provides a collection of marginal tables, i.e., the contingency tables for various subsets of the socioeconomic variables of interest. One must then attempt to reconstruct, as far as possible, the entire table from the available marginals. This problem is particularly important in forecasting the cell probabilities at a specified date, because the set of marginal tables available for updating and projection is usually much sparser than the set available for the base year. Typical sources are the U.S. Census (Fourth Count census tract data, Public Use Sample, and Urban Transportation Planning Package), metropolitan transportation surveys, screen line counts, and external forecasts of population and land use models. Local transportation surveys may provide observations from individual cells in the table at the survey date; other sources typically provide first and second order marginal distributions.

A classical method of combining contingency table data from two or more sources is iterative proportional fitting, associated with Deming and Stephan (1940). The method and its assumptions are discussed in Bishop, Fienberg, and Holland (1975). An earlier application to census data has been made by Liu (1976). The algorithm which we use is due to Haberman (1974).

For the illustrative contingency table of Figure 12, the procedure is as follows. Suppose we are given an initial trial table $p_{ij}^{(0)}$, and a set of observed marginals $\bar{p}_{i+}$ and $\bar{p}_{+j}$. Successive approximations are then given by

$$p_{ij}^{(n+1)} = p_{ij}^{(n)} \cdot \bar{p}_{i+} / p_{i+}^{(n)} \qquad (1)$$

and

$$p_{ij}^{(n+2)} = p_{ij}^{(n+1)} \cdot \bar{p}_{+j} / p_{+j}^{(n)} \tag{2}$$

(for $n = 0,2,4,...$), i.e., alternately rows and columns are rescaled to agree with the observed row and column sums. Under certain conditions (see Final Report Volume VIII, Appendix A) this iterative procedure always converges to a fitted table consistent with the given marginal data. A more general account of the algorithm and its properties is given in this reference, which also describes the interpretation of a contingency table in terms of "m-factor effects" by means of the log linear model.

One possible difficulty is that data may fail to include direct observations on interactions that are believed *a priori* to be important. Then, it may be desirable to attempt to recover the missing interactions by imposing sufficient structure on the data to identify these effects. In the absence of supplementary survey data, a case in point is the effect of socioeconomic variables on the workplace zone probabilities for a given zone of residence. To capture these interactions, we use the work destination model described in Part III, Chapter 2.

To combine data from different dates, we shall assume log cell probabilities follow a linear trend,[1]

$$\log p_\sigma(t) = A(t) + \alpha_\sigma(t - t_0) + \log p_\sigma(t_0) \quad , \tag{3}$$

where α_σ is the trend rate of change for the cell and $A(t)$ is a normalizing factor to satisfy $\sum_\sigma p_\sigma(t) = 1$. If t_0 and t_1 are two dates with observed data, then

[1] In the formulation of the contingency table in terms of the log linear model, this implies that the effects in the model with one or more factors exhibit linear trends. The fitting procedure implies that the order of effects exhibiting non-zero trends will not exceed the order of the highest order observed marginal table.

$$\alpha_\sigma = \frac{1}{t_1 - t_0} \log \frac{p_\sigma(t_1)}{p_\sigma(t_0)} - \frac{A(t_1)}{t_1 - t_0} \quad , \tag{4}$$

implying

$$\log\, p_\sigma(t) = \left\{ A(t) - A(t_1) \frac{t - t_0}{t_1 - t_0} \right\} + \frac{t - t_0}{t_1 - t_0} \log \frac{p_\sigma(t_1)}{p_\sigma(t_0)} + \log\, p_\sigma(t_0) \quad , \tag{5}$$

$$p_\sigma(t) = A'(t)\, p_\sigma(t_0) \left| \frac{p_\sigma(t_1)}{p_\sigma(t_0)} \right|^{\frac{t - t_0}{t_1 - t_0}} \quad . \tag{6}$$

In broad outline, our forecasting technique is to start from a common $\hat{p}_\sigma^{(0)}$ that reflects the interactions of all orders that appear to characterize the population in the geographical area under study. Typically, $\hat{p}_\sigma^{(0)}$ would be estimated from a sample of individual households, taken from a transportation survey, or, as in the present application, from the Census Public Use Sample. Then, iterative proportional fitting is applied to the observed marginals to refine the tables, first by residence zone and secondly by date. From the fitted tables $\hat{p}_\sigma(t)$ for various dates, cell trend rates are estimated. Using these trend rates, the fitted tables are extrapolated to the date at which a forecast is desired. This extrapolated table provides market segments and segment shares directly. Alternately, random or stratified sampling from the cells of the table provides a representative sample of a specified population. A further step is to associate with a sampled cell a case record of an observed household that appears in this cell. Such a record may contain added variables, or refinements of variables, which are not determined by the cell identification. Provided the household file from which this case record is drawn is representative, conditioned on cell identification, this method will provide a representative sample of the population.

The Iterative Proportional Fitting Method

The description of contingency tables

We first define the notation used to describe a multi-dimensional contingency table and its associated marginal tables. We recall the illustrative two-dimensional table of Figure 1, where the cell probabilities are denoted p_{ij} and the marginal probabilities are the row and column totals: $p_{i+} = \sum_j p_{ij}$ and $p_{+j} = \sum_j p_{ij}$. Generalizing this notation, assume K variables, indexed by the elements of $\underset{\sim}{K} = \{1,...,K\}$. Let I_k denote the set of categories defined for variable k . A cell in the contingency table is indexed by a vector $\sigma = (i_1,...,i_k)$; the set of possible cell indices is denoted $S = I_1 \times ... \times I_k$. Cell probabilities are denoted by p_σ . For any subset $B \subseteq S$, define $p_B = \sum_{\sigma \in B} p_\sigma$ to be the probability of B .

We wish to consider marginal probabilities for a subset of variables (or, configuration) $Q \subseteq \underset{\sim}{K}$; i.e., the probability of sets of the form

$$B = \underset{\ell \in Q}{X} \{i_\ell\} \times \underset{\ell \notin Q}{X} I_\ell \quad . \tag{7}$$

This probability can be denoted generally by p_B , but will also be denoted by an abbreviated notation: let σ be a K-vector with component k equal to i_k if $k \in Q$ and equal to "+" if $k \notin Q$, and let p_σ denote the probability of the set in equation (A.l) . For example, if $Q = \{1,2\}$ and $B = \{2\} \times \{3\} \times I_3 \times ... \times I_K$, then the probability of B is denoted $p_{23+...+}$.

One method of describing a contingency table is by use of a log linear model, in which the cell probabilities are written in the form

$$\log p_\sigma = \sum_{Q \subseteq \underset{\sim}{K}} u_{Q(\sigma)} \quad , \tag{8}$$

where $u_{Q(\sigma)}$ is a constant for each configuration Q and the cell in the Q-configuration marginal table that is the projection of σ . For example, if $K = 3$, then

$$\log p_{kj\ell} = u_{\varnothing} + u_{1(i)} + u_{2(j)} + u_{3(\ell)} + u_{12(ij)} + u_{23(j\ell)} + u_{13(i\ell)} + u_{123(ij\ell)} \quad , \tag{9}$$

where we write $u_{1(i)}$ rather than $u_{\{1\}(i)}$, etc. With the normalization that for any $k \in Q$,

$$\sum_{\sigma_k \in I_k} u_{Q(\sigma)} = 0 \quad , \tag{10}$$

this model has the same number of cells and independent effects $u_{Q(\sigma)}$, and can be inverted to express the effects in terms of the cell probabilities. Hence, any table can be described in terms of a log linear model representation. If Q has m elements, then $u_{Q(\sigma)}$ is referred to as an m-factor effect or an effect of order m.

If all the variables in a contingency table are independent, then all the effects u_Q will be zero except the zero and one factor effects. Then the table can be reconstructed from its first order marginals. More generally, if a table has effects of order m or higher equal to zero, then it can be reconstructed from the family of $(m - 1)$ order marginals.

Properties of the IPF method

The iterative proportional fitting procedure appears to have been first discussed by Deming and Stephan (1940), and is treated in some detail in Bishop, Fienberg and Holland (1975). Other useful references are Darroch (1962) and Haberman (1974). The method enables one to adjust a contingency table so as to be consistent with an observed set of marginal tables.

The iterative proportional fitting algorithm applied to data available at a common date has a simple description. Suppose that observations are available on the probabilities of a sequence of marginal distributions with configuration $Q_1,...,Q_J$. We assume each of these configurations to be maximal in the sense that none are contained in another. For each $\sigma \in S$, let σ_j denote the index vector formed by replacing the components k of σ for which $k \notin Q_j$ by "+" . Then, σ_j indexes the cell of the marginal distribution with configuration Q_j that is the projection of σ . For example, in Figure 1, if $Q_2 = \{1\}$ is the configuration giving the first order marginal distribution of residence zone and $\sigma = (4,3)$, then $\sigma_2 = (4,+)$. Suppose an initial trial table $\hat{p}_\sigma^{(0)}$ for $\sigma \in S$ is given. Then the table is modified iteratively using the formula

$$(11) \qquad \hat{p}_\sigma^{(i+1)} = \hat{p}_\sigma^{(i)} \frac{\bar{p}_{\sigma_j}}{\hat{p}_{\sigma_j}(i)} \quad , \quad (\sigma \in S)$$

where j cycles through the values $j = 1,...,J$ in successive iterations. If the observed marginal distributions are mutually consistent[1] and the initial trial values $\hat{p}_\sigma^{(0)}$ are positive, then this algorithm always converges to a fitted table $\hat{p}_\sigma$ that is consistent with the observed marginal distributions; i.e., $\hat{p}_{\sigma_j} = \bar{p}_{\sigma_j}$ for $\sigma \in S$ and $j = 1,...,J$ (Haberman, 1974).

The iterative proportional fitting algorithm has the following properties (Bishop, Fienberg and Holland (1975)):

1. If the initial trial table $\hat{p}_\sigma^{(0)}$ has no non-zero effects of order greater than the order m of the largest observed marginal configuration, then

[1]The conditions for marginal distributions to be mutually consistent have been given by Darroch (1962). A necessary condition is that their common marginals agree. For example, the marginal configurations $Q_1 = \{1,2,3\}$ and $Q_2 = \{1,2,4\}$ must have identical marginals for the configuration $Q_3 = \{1,2\}$.

the final fitted table $\hat{p}_\sigma$ will have no non-zero effects of order greater than m . (In particular, if $\hat{p}_\sigma^{(0)}$ is the same for all σ , this conclusion holds.)

2. If the initial trial table $\hat{p}_\sigma^{(0)}$ has non-zero effects of order greater than the order m of the largest observed marginal configuration, then all effects in the final fitted table $\hat{p}_\sigma$ of order greater than m will equal the corresponding effects in $\hat{p}_\sigma^{(0)}$.

3. The final fitted table $\hat{p}_\sigma$ gives the unique maximum likelihood estimate of the log linear model, subject to the condition that $\hat{p}_\sigma$ and $\hat{p}_\sigma^{(0)}$ have the same effects for orders exceeding the highest order marginal.