Institute of Transportation Studies
109 McLaughlin Hall
University of California
Berkeley, California 94720

THE URBAN TRAVEL DEMAND FORECASTING PROJECT

PHASE I FINAL REPORT SERIES

VOLUME VIII

# DEMOGRAPHIC DATA FOR POLICY ANALYSIS

by

Daniel McFadden
Stephen Cosslett
Gerald Duguay
Woo Sik Jung

June 1977

Research was supported in part by the National Science Foundation, through grants GI-43740 and APR74-20392, Research Applied to National Needs Program, and by the Alfred P. Sloan Foundation, through grant 74-12-8, to the University of California, Berkeley.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The results would not have been possible without the helpful provision of source data by the Metropolitan Transportation Commission. The generous financial support from the National Science Foundation and the Alfred P. Sloan Foundation is gratefully acknowledged. The principal researcher is finally responsible for the contents of this report.

# DEMOGRAPHIC DATA FOR POLICY ANALYSIS

## 1. 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.

SYNSAM is a methodology for generating a synthetic representative sample of households for an urban area for any specified date. 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 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. Since such data are available for all Standard Metropolitan Statistical Areas (SMSA), the procedure is readily transferable to other cities.

We also provide a description of the software that was used to implement the two principal computational steps in the procedure, and which is available as a set of FORTRAN routines. These are sufficiently flexible to allow application to other geographic areas and to different sets of socioeconomic variables. Options include the construction of samples weighted according to residence zone.

A principal feature of the SYNSAM procedure is the use of Iterative Proportional Fitting (IPF) to construct and update the contingency table for zone of residence and for 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). An account of this program, together with notes on its use, is given in Appendix C. The other principal step is to actually construct the 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. Some programming notes are given in Appendix D which may be taken as a "user's guide" in the event that the program is used to synthesize other samples for market segments or for different areas.

Since SYNSAM is intended to be a flexible methodology, capable of accommodating a variety of data sources and estimates, we discuss also alternative methods and possible improvements in some steps of the procedure.

## 2. 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 1. 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 1, cell probabilities are denoted $p_{ij}$, where $i$ is the level of the first variable, and $j$ is the level of the second. The marginal probabilities are denoted $p_{i+} = p_{i1} + p_{i2} + p_{i3}$ for the number of persons in household and $p_{+j} = p_{1j} + p_{2j} + p_{3j} + p_{4j} + p_{5j}$

FIGURE 1. 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 | $p_{+1}$ | $p_{+2}$ | $p_{+3}$ | $p_{++}=1$ |

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 2. Notation suitable for treatment of the general case is defined in Appendix A .

Cell probabilities are denoted by $p_\sigma$ , where the index is a vector $\sigma = (i_1,\ldots,i_9)$ 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, since 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 1, 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

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

and

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

(for $n = 0,2,4,\ldots$), i.e., alternately rows and columns are rescaled to agree with the observed row and column sums. Under certain conditions (see 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 Appendix A, which also describes the interpretation of a contingency table in terms of "m-factor effects" by means of the log linear model. In addition, Appendix A discusses practical expedients that may be used when the number of cells in the table becomes too large for a straightforward application of the IPF algorithm.

One possible difficulty is that data may fail to include direct observations on interactions which 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 combine data from different dates, we shall assume log cell probabilities follow a linear trend,

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

where $\alpha_\sigma$ is the trend rate of change for the cell and $A(t)$ is a normalizing factor to satisfy $\sum_\sigma p_\sigma(t) = 1$ .

In broad outline, our forecasting technique is to start from a common $\hat{p}_\sigma^{(0)}$ which reflects the interactions of all orders which 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 which 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.

# 3. Application of SYNSAM to the San Francisco Bay Area

## 3.1 Orientation

The Short Run Generalized Policy (SRGP) disaggregate model system developed by Cambridge Systematics, Inc., for the Metropolitan Transportation Commission, along with model modifications suggested by the regional policy analysis case study of the Urban Travel Demand Forecasting Project, provides the starting point for this sample synthesis. This model system forecasts auto ownership, work trip mode, and non-work trip generation, distribution, and mode, taking as exogenous population, residence and work location, and socioeconomic characteristics. The desired exogenous variable forecasts are listed in Table 1. Note, however, that the present application of the SYNSAM procedure cannot provide "person" variables (as opposed to "household" variables) that may be correlated with zone of employment: this applies to items 9 (type of employment) and 10 (hourly wage) in Table 1.

The nine socioeconomic variables used to classify households (besides residence and employment zones) are given in Table 2. The number of socioeconomic cells defined by these variables, i.e., the number of possible types of household in this classification scheme, is then 2.304. Not all these possible cases actually occur: approximately half the cells are found to be empty in 1/100 Census data for area, mainly because 1.152 cells belong to the first category of the 7th variable. Table 2 also gives the correspondence between these nine variables and the tabulations in the Public Use Sample (SMSA 15%) of the 1970 Census.

TABLE 1. Exogenous Variables in the SRGP Disaggregate Model System

1. Family income*
2. Number of household workers*
3. Age of head of household
4. Number of drivers in household
5. Number of children
6. Number of adults
7. Length of residence
8. Number of household autos (needed only when auto ownership portion of model system is not used)
9. Type of employment
10. Hourly wage
11. Residence location zone*
12. Work location zone*
13. Race

*Essential items.

TABLE 2. Socioeconomic Variables Used in the SYNSAM Classification of Households

| Variable | Symbol | Categories | 1970 PUS Tabulations |
|---|---|---|---|
| 1. Workers | W | 1. Zero workers<br>2. One worker<br>3. Two or more workers | P31 |
| 2. Family type | F | 1. Husband and wife, head under 45<br>2. Husband and wife, head over 45<br>3. Other family<br>4. Primary individual | H70 and H72-73 |
| 3. Autos | A | 1. Zero autos<br>2. One auto<br>3. Two or more autos | H60 |
| 4. Income | I | 1. Above Bay Area median<br>2. Otherwise | H85-87<br>($10,500 in 1970) |
| 5. Persons | P | 1. One or two person household<br>2. Otherwise | H12-13 |
| 6. Units | U | 1. One unit attached or detached<br>2. Otherwise | H35 |
| 7. Race | B | 1. Black<br>2. Otherwise | H71 |
| 8. Tenure | R | 1. Renter occupied<br>2. Otherwise | H31-33 |
| 9. Mobility | M | 1. Head moved in past five years<br>2. Otherwise | H90 |

For definiteness, the numbers quoted below apply to the sample for the nine-county San Francisco Bay Area, although the procedure is, of course, more generally applicable. The basic geographic unit is the traffic analysis zone (TAZ), of which there are 440. These will be referred to as "zones." The area is subdivided also into 1.058 census tracts in the 1970 Census.

## 3.2 Data Sources

The data sources available for the San Francisco Bay Are are the following.

1. Public Use Sample (PUS), 1970 Census, a file of individual household records, identified by county of residence, with approximately 14,000 households (one-fifteenth of the 15% census sample) in the San Francisco-Oakland and San Jose SMSA's. This sample includes all socioeconomic variables of interest, including number of workers per household, but does not identify residence or work zone.

2. Fourth Count Census of Population and Housing, 1970, a file of census tract counts containing all socioeconomic variables of interest, and destination counts for selected destination districts, reported in one, two, and three-way marginal tables.

3. Second Count Census of Housing, 1960 and 1970, a file of county summary data for the same marginals as the 1970 Fourth Count.

4. Urban Transportation Planning Package, 1970 Census, a file of traffic analysis zone counts giving marginal tables on all the socio-economic variables of interest, and tables of work trip flows zone to zone. No marginals are given for interactions of socioeconomic variables

and origin to destination flows.

5. PLUM model (*) , a land use model giving total population, number of households, number of employed residents, and employment projections for each traffic analysis zone. Published projections are available for 1980, and rates of growth can be inferred for the decade 1970-1980.

In addition to these sources, a potentially very useful set of data is available in BATSC, a transportation survey of 30,521 households taken in 1965 and including household records of almost all variables of interest. The BATSC sample was designed to be random, but reports from individuals involved in the data collection and processing, and comparison of marginals with external sources, indicate major departures from randomness. Details of this comparison are given in Appendix B.

These difficulties with the representativeness of the BATSC sample make it a poor foundation for synthesis of a representative sample. Consequently, we have utilized only Census data in the sample synthesis. This has the disadvantage of providing inadequate data on some interactions which are believed a priori to be important, such as interactions between socioeconomic characteristics and work destination zone. However, there is an advantage in that the method is transferable to other metropolitan areas.

---

(*) Population and Land Use Model, developed by ABAG (Association of Bay Area Governments)

## 3.3 Outline of the Procedure

There are four main stages in the synthesis, which are described in detail in the following sections.

Stage I: the contingency table for all socioeconomic variables and residence zone is computed for the required year. This table is obtained in the form of a set of 2.304-cell socioeconomic contingency tables, one for each zone of residence, plus a set of population projections by zone.

Stage II: a set of coefficients is estimated, which allows the contingency table to be extended to destination zone probabilitie for any given socioeconomic cell and zone of residence.

Stage III: cases are synthesized by random sampling from the tables constructed in stages I and II. To each synthesized case are attached the household and person records of a PUS case in the same socioeconomic cell.

Stage IV: certain quantitative variables in the PUS records are updated from 1970 to the required year. The only such variable of interest here is family income. (The updated contingency tables give only income above or below the contemporary median, rather than a dollar amount.) This involves a separate econometric study of the growth rates of wages and incomes in the Bay Area.

## 4. Stage I: Iterative Proportional Fitting

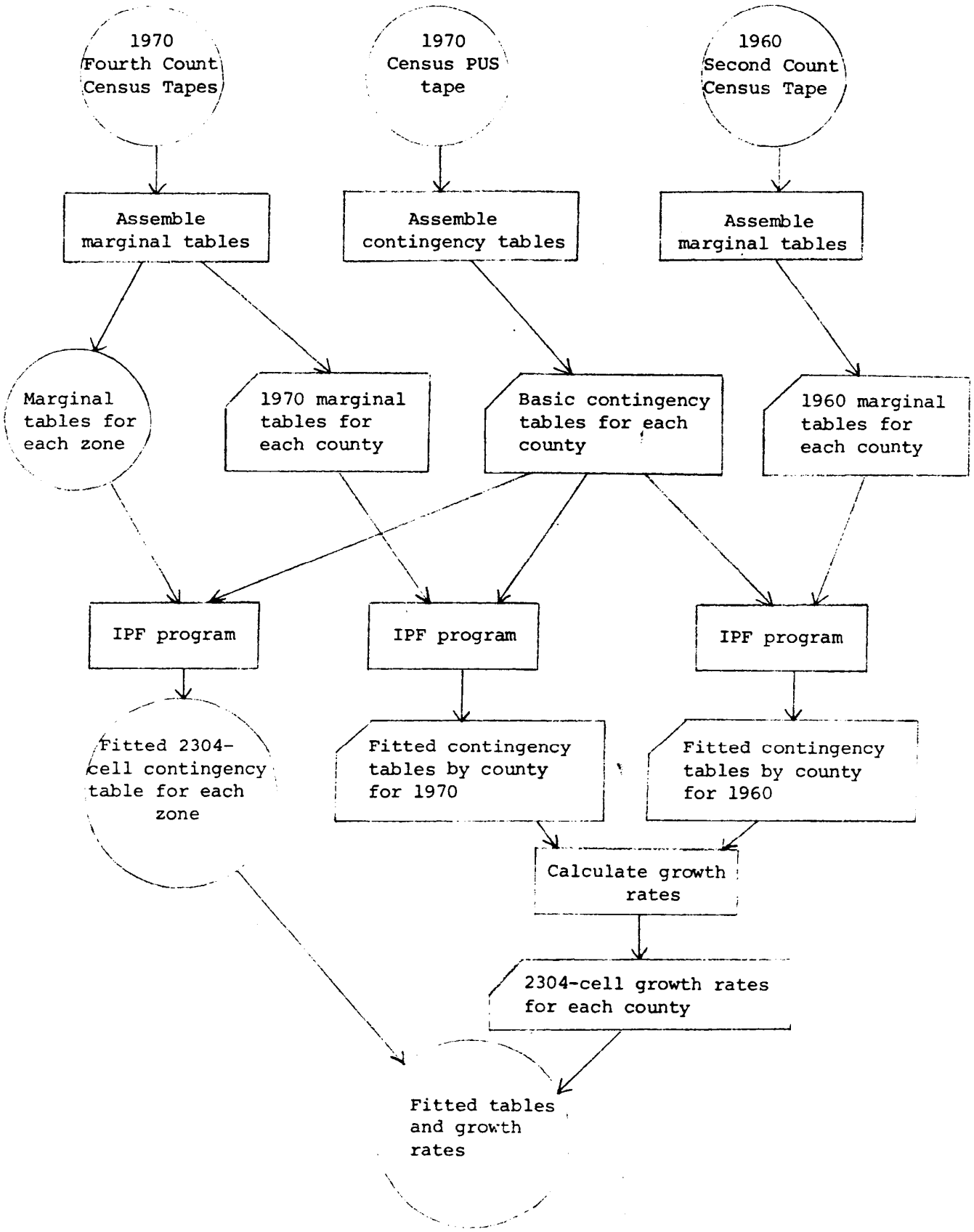
For each zone of residence, one requires a sampling table which gives the probabilities of a household falling into each of the 2.304 socioeconomic cells defined by Table 2. Construction of these tables incurs the major part of the computing costs.

These sampling tables cannot be obtained directly from census data, for the following reason. On the one hand, the 1970 PUS gives all the socioeconomic variables of interest for each household in a 1/100 census sample, so that each PUS case can be assigned to one of the 2.304 cells; but residence is given only by county rather than by zone or tract (to ensure confidentiality). On the other hand, the 1970 Fourth Count of Population and Housing gives counts by census tract for all the socioeconomic variables of interest, but reports them only in one-, two-, and three-way marginal tables rather than the nine-way tables required.

Iterative proportional fitting is therefore used to construct a full contingency table consistent with the marginal data. The initial approximation is the table of cell probabilities for a given county, obtained from the 1970 PUS. The IPF procedure is then used in two ways: first, to correct for geographic variation in cell probabilities between different zones in the county; and secondly to correct cell probabilities for secular variation.

Sampling tables were constructed as follows.

FIGURE 2. Stage I: Estimation of Socioeconomic Contingency Tables by Iterative Proportional Fitting



(1) A 2.304-cell socioeconomic table was formed from the 1970 PUS data for each county.

(2) The twenty marginal contingency tables listed in Table 3 were used. These tables from the 1970 Census Fourth Count were aggregated from census tracts to the 440 zone level.

(3) For each zone, an iterative proportional fit was carried out, starting from the 1970 PUS table for the county and using the twenty marginals for the zone from the Fourth Count.

These three steps result in a 2.304-cell table for 1970 for each of the 440 zones. Next, trend forecasts were obtained by extrapolating from the growth rates between 1960 and 1970, as follows.

(1) For each county, an iterative proportional fit was carried out for 1960, starting from the 1970 PUS table. The marginal data were obtained from the Second Count Census of Housing for 1960. These counts give the same twenty marginal tables, listed in Table 3, but aggregated from tracts to the county level.

(2) For each county, an IPF was carried out for 1970, starting from the 1970 PUS table and using marginal data for 1970. Let $p_\sigma^c(1960)$ and $p_\sigma^c(1970)$ denote the fitted cell probabilities for cell $\sigma$ and county $c$ for these dates.

(3) The annual rate of change $\rho_\sigma^c$ of the cell probability $p_\sigma^c$ is then given by

$$\rho_\sigma^c = [p_\sigma^c(1970)/p_\sigma^c(1960)]^{1/10} \quad . \tag{4}$$

TABLE 3. Marginal Socioeconomic Tables

| Variables | Number of cells | 1970 Fourth Count tabulations used |
|---|---|---|
| 1. Workers | 3 | (*) |
| 2. Race × Tenure × Units | 8 | H9 |
| 3. Race × Tenure × Mobility | 8 | H10 |
| 4. Race × Tenure × Persons | 8 | H37 |
| 5. Family type × Mobility × Race | 16 | H109 |
| 6. Family type × Mobility × Tenure | 16 | H109 |
| 7. Family type × Income × Race | 16 | H110 |
| 8. Family type × Income × Tenure | 16 | H110 |
| 9. Family type × Units × Race | 16 | H111 & H112 |
| 10. Family type × Units × Tenure | 16 | H111 & H112 |
| 11. Persons × Units × Race | 8 | H114 & H115 |
| 12. Persons × Units × Tenure | 8 | H114 & H115 |
| 13. Persons × Mobility × Race | 8 | H116 |
| 14. Persons × Mobility × Tenure | 8 | H116 |
| 15. Persons × Income × Race | 8 | H117 |
| 16. Persons × Income × Tenure | 8 | H117 |
| 17. Units × Mobility × Race | 8 | H138 & H139 |
| 18. Units × Mobility × Tenure | 8 | H138 & H139 |
| 19. Units × Autos × Race | 12 | H140 & H141 |
| 20. Units × Autos × Tenure | 12 | H140 & H141 |

(*) Available, at the zone level, from the 1970 Census UTPP file (Tabulation IC4).

(4) The 2.304-cell table for zone $z$ in year $t$ can then be forecast to be

$$(5) \qquad p_{\sigma}^{z}(t) = A_{z}(t) \cdot p_{\sigma}^{z}(1970) \cdot [\rho_{\sigma}^{c}]^{(t-1970)}$$

where $c$ is the county containing zone $z$, and $A_z(t)$ is set so the probabilities in zone $z$ sum to one.

This method of obtaining cell probabilities for 1960 is necessary because (a) the 1960 census marginal data is not available in machine-readable form for individual tracts, but only as county summaries, and (b) PUS data for 1960 is not available at the county level. Because of (a), it is not feasible to take into account differences in trends between individual zones within a county. In view of (b), it is not possible to determine any trend in effects of fourth order or higher in the contingency tables, because the same initial approximation is used for both dates and the updating marginal tables are at most three-way tables (see Appendix A for an account of properties of the IPF algorithm).

Although we refer to "counties" in connection with these tables, the Public Use Sample actually identifies the following <u>five</u> geographical divisions in the San Francisco Bay Area.

1. San Francisco
2. Alameda County
3. Contra Costa and Marin Counties
4. San Mateo County
5. San Jose SMSA

These cover a somewhat smaller area than the nine-county area covered by the 440 TAZ's . For the remaining counties (Napa, Solano, and Sonoma), the initial cell probabilities for the IPF procedure were taken from Contra Costa and Marin counties, as were the growth rates $\rho_{\sigma}^{c}$ .

Output from Stage I consists of: (a) a 2.304-cell socioeconomic sampling table for each of the 440 residence zones for 1970; and (b) a table of 2.304 growth rates $\rho_{\sigma}^{c}$ for each of the five "counties" of residence, defined above.

## 5. Stage II: Destination District Probabilities

### 5.1 The Method

The next step is to extend the sampling table, for each origin zone, to include zone of employment. First note that the magnitude of each table, $2.304 \times 440 = 1{,}013{,}760$ cells, makes it impractical to estimate or operate on the whole table at once. Second, we note that none of the census data sources provide marginal configurations which yield direct information on interactions between socioeconomic variables and destinations. (The 1970 PUS provides a relatively uninformative classification of work location: in CBD, central city, ring, or outside SMSA.)

In view of these facts, we have imposed a structure on the data which permits identification of socioeconomic-destination interactions, and have adopted a sampling procedure which avoids computation of the complete table. As a result, a comparatively small number of estimated coefficients will then allow computation of the required destination zone probabilities for any given origin zone and socioeconomic cell.

The first step is to aggregate the zones into larger units which we refer to as districts. The six origin districts and thirteen destination districts used in the present application are defined in Table 4. For a given origin district, the destination district probabilities will then be estimated as functions of socioeconomic variables in a linear probability model. The household socioeconomic variables of Table 2 are expressed in terms of the set of zero-one variables (or "dummy variables") defined in Table 5.

FIGURE 3. Stage II: Estimation of Linear Probability Model for Destination Districts

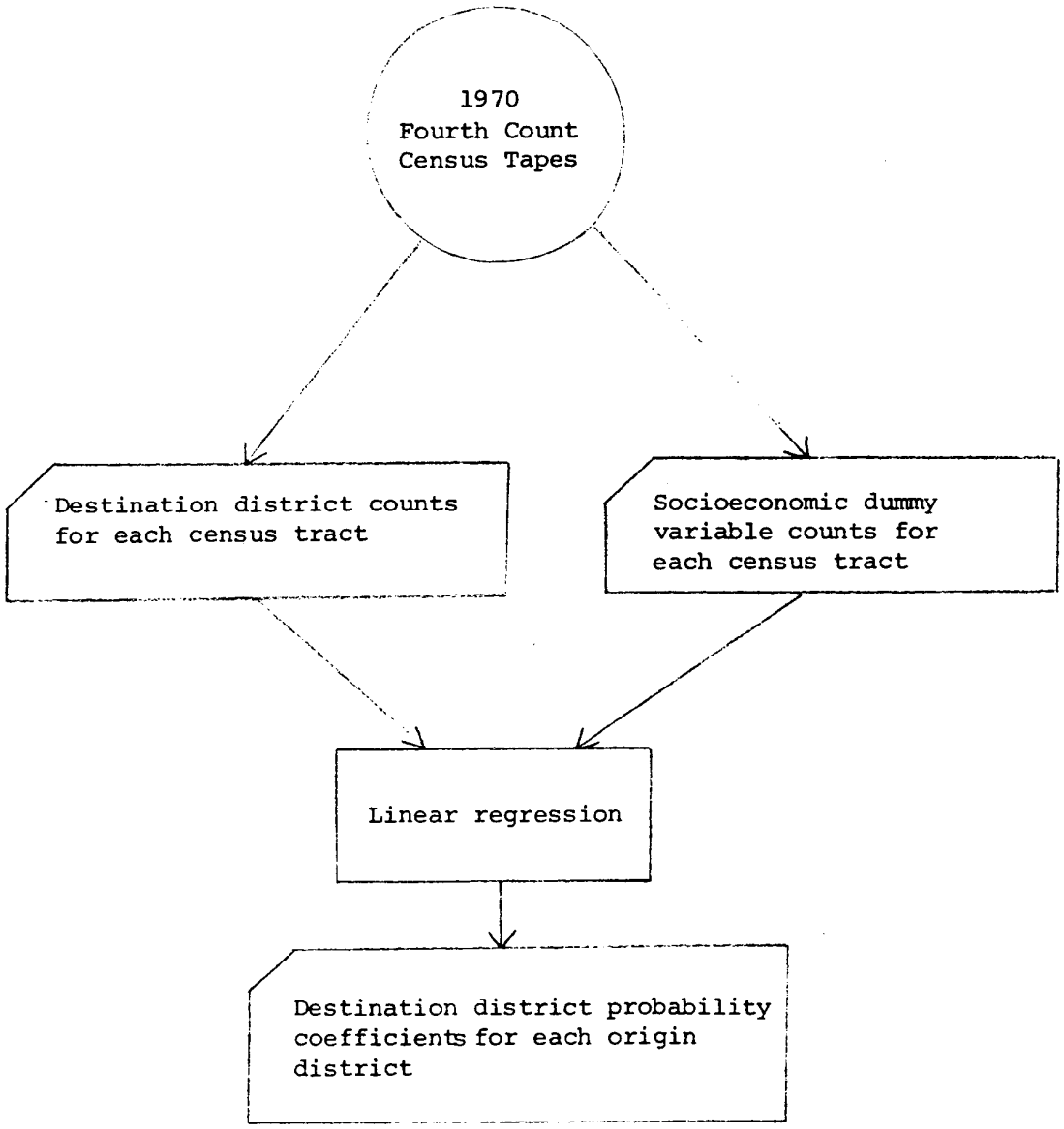


TABLE 4. Definition of Origin and Destination Districts

Origin Districts

1. Marin and Sonoma Counties
2. Napa, Solano, and Contra Costa Counties
3. Alameda County
4. Santa Clara County
5. San Mateo County
6. San Francisco County

Destination Districts

1. San Francisco CBD (Central Business District)
2. San Francisco, excluding CBD
3. Oakland CBD
4. Oakland, excluding CBD
5. Alameda County, excluding Oakland
6. San Mateo County
7. Contra Costa County
8. Marin County
9. San Jose City
10. Santa Clara County, excluding San Jose
11. Sonoma County
12. Solano County
13. Napa County
14. Destinations outside San Francisco Bay Area

TABLE 5. Definition of Variables in the Linear Probability Model for Destination Districts

Each variable equals one if the associated condition is true; otherwise it equals zero. The variables W, F, A, etc., appearing in the conditions, are the socioeconomic variables defined in Table 2.

| Variable | Condition |
|---|---|
| $v^{(1)}$ | (always equals 1) |
| $v^{(2)}$ | W = 1 |
| $v^{(3)}$ | W = 2 |
| $v^{(4)}$ | F = 1 |
| $v^{(5)}$ | F = 2 |
| $v^{(6)}$ | F = 3 |
| $v^{(7)}$ | A = 1 |
| $v^{(8)}$ | A = 2 |
| $v^{(9)}$ | I = 1 |
| $v^{(10)}$ | P = 1 |
| $v^{(11)}$ | U = 1 |
| $v^{(12)}$ | B = 1 |
| $v^{(13)}$ | R = 1 |
| $v^{(14)}$ | M = 1 |

Assume that the probability of destination district $\delta$, conditioned on census tract $\tau$ in origin district $d$ and on socioeconomic cell $\sigma$, is a linear function of the corresponding zero-one variables $v_\sigma^{(k)}$ $(k = 1,\ldots,14)$, i.e.,

$$(6) \qquad p_{\delta|\tau\sigma} = b_{d\delta}^{(1)} v_\sigma^{(1)} + b_{d\delta}^{(2)} v_\sigma^{(2)} + \ldots + b_{d\delta}^{(14)} v_\sigma^{(14)} + \varepsilon_{\tau\sigma} \ .$$

The coefficients $b^{(k)}$ are assumed independent of census tract within a given origin district. Aggregating over households in census tract $\tau$ in district $d$, one obtains

$$(7) \qquad p_{\delta|\tau} = b_{d\delta}^{(1)} \bar{v}_\tau^{(1)} + b_{d\delta}^{(2)} \bar{v}_\tau^{(2)} + \ldots + b_{d\delta}^{(14)} \bar{v}_\tau^{(14)} + \varepsilon_\tau \ ,$$

where

$p_{\delta|\tau}$ = the proportion of households in census tract $\tau$ with work destinations in district $\delta$,

and

$\bar{v}_\tau^{(k)}$ = the average value of $v_\sigma^{(k)}$ in census tract $\tau$

= the proportion of households in census tract $\tau$ for which $v_\sigma^{(k)} = 1$.

Least squares applied to equation (7) for each origin and destination district, taking as observations the census tracts in the origin district and weighting each observation by the number of workers in the tract, allows estimation of the coefficients $b_{d\delta}^{(k)}$ for each origin district. With six origin districts and fourteen destination districts, this requires eighty-four regressions, of which six are redundant and

provide checks on the analysis. There are from ninety-three to 278 census tracts in each origin district.

The estimated coefficients $b_{d\delta}^{(k)}$, when substituted in equation (6), allow forecasts of the probability of a destination $\delta$, conditioned on socioeconomic cell $\sigma$ and a given origin district. The linear probability model structure does not guarantee that all these computed probabilities are nonnegative. Hence, an ad hoc procedure is followed for any selected socioeconomic cell $\sigma$: the probability $p_{\delta|d\sigma}$ is redefined to equal

$$(8) \qquad p_{\delta|d\sigma} = \frac{\text{Max } (0, p'_{\delta|d\sigma})}{\sum_{\delta=1}^{12} \text{Max } (0, p'_{\delta|d\sigma})} ,$$

where $p'_{\delta|d\sigma}$ is the value given by equation (6) for an origin district $d$.

Given the socioeconomic cell probabilities $p_\sigma$ and the estimated conditional destination probabilities $p_{\delta|d\sigma}$, we can in principle provide a complete table $p_{\delta|z\sigma}p_{\sigma|z}$ of the socioeconomic characteristics and destinations for households in zone $z$ (where $d$ is the origin district containing zone $z$). Because of the large size of this table, we shall ordinarily compute cell probabilities only for selected cells.

The procedure above provides a destination district, but not a destination zone. To provide this last step, we utilize the marginal distribution of destination zones from each origin zone which is available from the 1970 UTPP Census file. We assume that <u>conditional on destination district</u>, the distribution of destination zones is independent of socioeconomic characteristics, i.e., if $p_{y|\delta z\sigma}$ denotes the probability of destination zone $y$, conditioned on destination district $\delta$, origin zone $z$, and socioeconomic cell $\sigma$, then $p_{y|\delta z\sigma} = p_{y|\delta z}$. But $p_{y|z}$ is given by UTPP, and

$$(9) \qquad p_{y|\delta z} = p_{y|z} \Big/ \sum_{y \in D_\delta} p_{y|z} \quad ,$$

where $D_\delta$ is the set of zones in destination district $\delta$. Then, $p_{y|\delta z} p_{\delta|z\sigma} p_{\sigma|z}$ is the cell probability in the grand table for origin zone $z$, destination zone $y$, and socioeconomic characteristics $\sigma$.

## 5.2 Discussion of the Estimation Procedure

The linear model of equation (7) can be estimated from data at the census tract level given by the 1970 Census Fourth Count of Population and Housing. The independent variables $\bar{v}_\tau^{(k)}$ are obtained from the same tabulations as the marginal tables used in the IPF procedure of Stage I. A tabulation is also available of counts by district of employment for each tract of residence, which provides the dependent variables $p_{\delta|\tau}$. There are, however, two slight complications to consider. First, the variable $W$ is not directly available at the tract level: the proportions of households in the tract with zero, one, and two or more members at work have to be estimated indirectly, using the average numbers of employed persons per household in the tract, the unemployment rate in the tract, and the number of households whose demographic characteristics indicate that they have no member in the labor force. Secondly, it is necessary to use destination districts as defined by the Census Bureau, which might in some circumstances be found too large, and which are defined in different ways according to the county of residence.

An alternative procedure, which was employed in the present case, is to take zones rather than tracts as the observations. The destination district probabilities $p_{\delta|z}$ may then be obtained by aggregating the

zone-to-zone trip counts of the 1970 Census UTPP file,[1] using any desired definition of origin and destination districts in terms of zones. There is also the advantage that, at the zone level, tabulation of households by number of workers is given by UTPP. The disadvantage, however, is that the origin districts have to be relatively large so as to ensure an adequate number of observations in each regression.

We should note that for some origin districts, particularly Alameda County, the residuals from the linear regression are undesirably large: for certain individual zones, the fitted destination district probabilities differ by as much as 30% from their actual values (although, of course, such discrepancies disappear when origin zones are aggregated into districts). This suggests these origin districts are too coarsely drawn and thus are internally inhomogeneous, with geographic location playing a major role, independently of the socioeconomic variables, in determining destination district probabilities. This problem may be exacerbated here by the presence of three urban centers in the area (San Francisco, Oakland, and San Jose). A change to smaller origin districts may require the use of individual census tracts, rather than zones, as observations. The effects of using redefined origin districts, based on the pattern of flows to destination districts rather than on administrative boundaries, are currently under investigation.

---

[1]There are differences of the order of 5-10% between aggregate destination district probabilities obtained from the Fourth Count and from UTPP (mainly due to allocation of unidentified work locations). We used an improved version of the UTPP work trip tables kindly made available by Mr. Pat Hackett of the Metropolitan Transportation Commission.

One may also utilize supplementary data that can provide more direct information on the interaction between socioeconomic variables and zone of employment. This would also permit the use of person (as opposed to household) variables: for example, whether or not the worker is a head of household. The BATSC sample may be used to estimate these interactions, for a reduced set of socioeconomic variables, provided that the sampling is random within each socioeconomic cell (even though, as mentioned above, the sampling over cells appears significantly non-random). However, the procedure would then no longer be readily transferable to other metropolitan areas.

The method does not currently allow for trends in work destination probabilities, assuming these to remain at 1970 levels. This assumption could be relaxed by taking PLUM forecasts of TAZ workers in residence and employment as updated marginals for the UTPP origin-destination table, assuming higher order interactions remain unchanged, and applying iterative proportional fitting to the UTPP table. The size of this table (193,600 cells) makes this computation cumbersome; however, convergence should be rapid. Screen line counts can also be interpreted as marginals to this table which can be incorporated in the iterative proportional fit of the trip table. We note that substantial economies in fitting could be obtained by adjusting the table giving district to district probabilities, and assuming distribution conditioned on district is stationary.

## 6. Stage III: Construction of the Synthetic Sample

### 6.1 Sampling Tables and Other Input Data

A file of household case records, consistent with the contingency table computed in Stages I and II, is now constructed. We describe a method which permits the selection of random or stratified samples, as required for policy applications, along with sampling weights, for any specified year. This procedure has been implemented by a computer program, which is described in more detail in Appendix D.

First, we list the inputs to the procedure. The actual sequence and format of these inputs are also given in Appendix D.

(1) Data on the number of households in each zone. From PLUM data from 1970 and projections for 1980, we obtain: (a) the number of households in each zone in the base year 1970; and (b) the projected ten-year growth rate for each zone between 1970 and 1980.

(2) The socioeconomic sampling tables computed in Stage I: (a) the 2.304 probabilities $p_{\sigma}^{z}(1970)$ for each zone of residence $z$ ; and (b) the 2.304 annual growth rates $\rho_{\sigma}^{c}$ for each county of residence $c$ .

(3) The destination district probability coefficients $b_{d\delta}^{(k)}$ $(k = 1,2,\ldots,14)$ estimated in Stage II for each origin district $d$ and destination district $\delta$ .

(4) The zone-to-zone trip tables from the 1970 Census UTPP file, which tabulates for each residence zone the total number of work trips

FIGURE 4. Stage III: Construction of Synthetic Sample

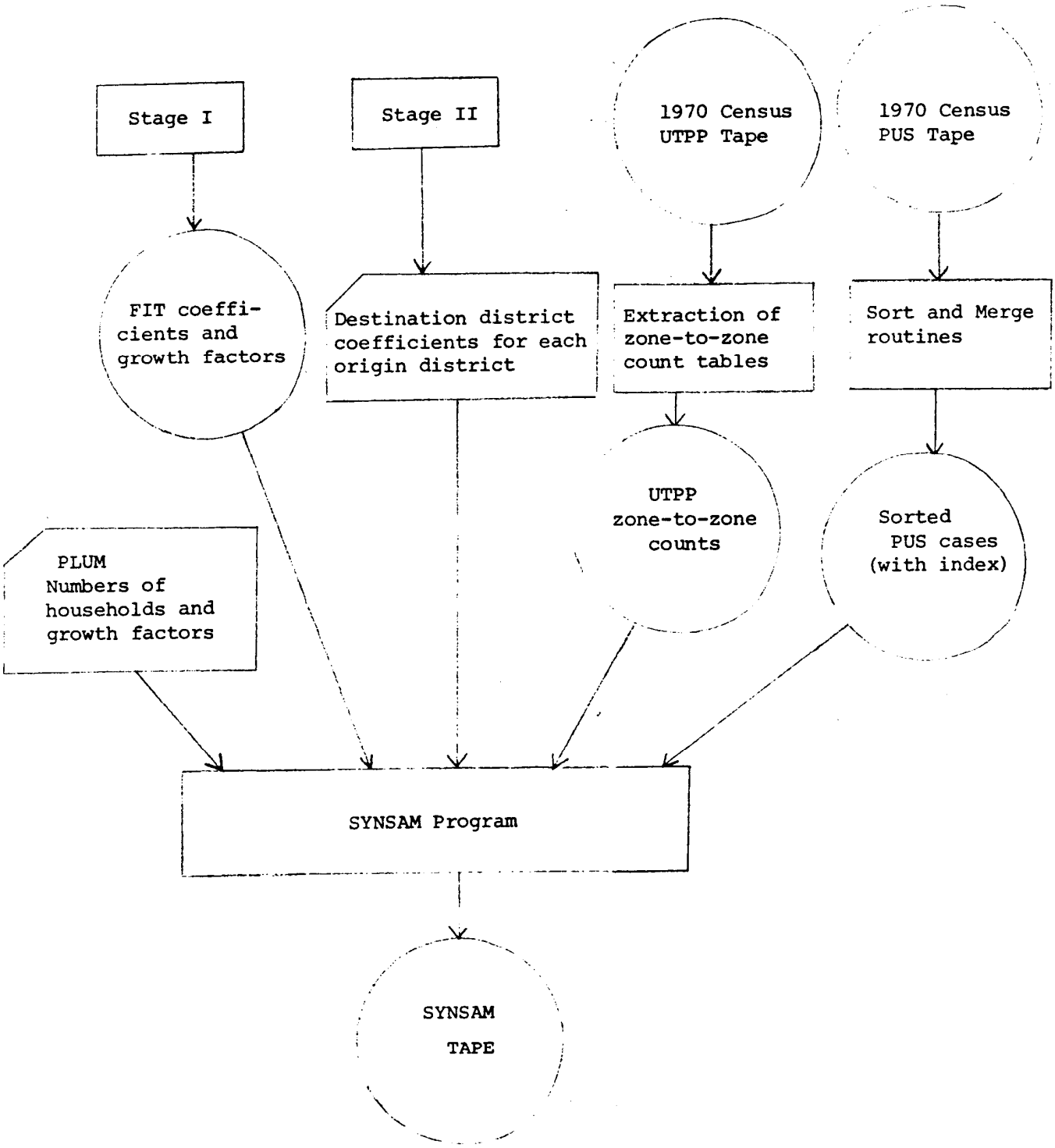
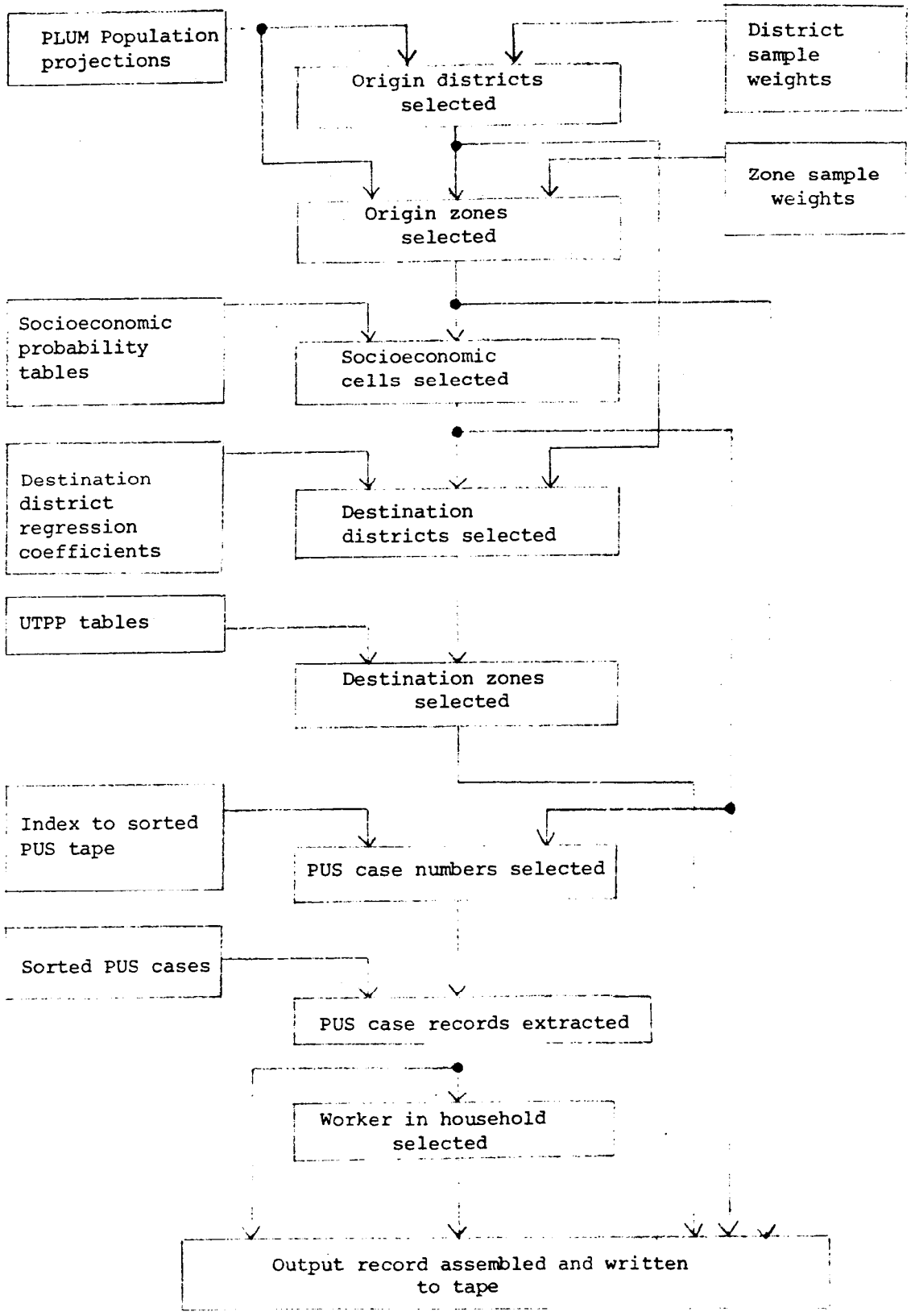


FIGURE 5. Sampling Procedure: Outline of Data Flow



to each zone of employment.

(5) A file of all 1970 Census PUS cases in the San Francisco-Oakland and San Jose SMSA's (but excluding vacant housing units and individuals living in group quarters). In order to expedite the sampling procedure, this file was first sorted according to socio-economic cell $\sigma$, and an index table was constructed giving the number of PUS cases in each of the 2.304 cells.

## 6.2 Sampling Procedure

We construct our sample on the assumption that a household is sampled randomly (according to a specified stratification), and that an individual is sampled randomly from the employed adults in the household. Therefore, a key forecast is number of households. Let $N$ be the number of synthesized cases required, and let $t$ be the desired forecast year.

(1) From the PLUM data, compute $n_z$, the projected number of households in each zone in year $t$. Aggregate these to give $n_d$, the number of households in each origin district $d$, and also $n$, the total number of households in the area. The corresponding origin zone and origin district probabilities are $\pi_z = n_z/n$ and $\pi_d = n_d/n$. The value of $n$ is subsequently used to obtain the expansion factor from the sample to the entire area population.

(2) Distribute the $N$ households among origin districts, by random sampling with probabilities $r_d$. The district sampling weights

are then $\pi_d/r_d$ . The sampling probabilities $r_d$ are, in general, specified by the user; in the default case the $r_d$ are taken equal to the actual probabilities $\pi_d$ . Let $N_d$ be the number of cases assigned to origin district $d$ .

(3) Distribute the $N_d$ households in each origin district among the constituent origin zones, by random sampling with probabilities $r_{z|d}$ . The origin zone sampling weight is then $\pi_{z|d}/r_{z|d}$ , where

$$\pi_{z|d} = \pi_z \Big/ \sum_{z \in D_d} \pi_z \quad . \tag{10}$$

$D_d$ is the set of zones in district $d$ .

(4) Update the probabilities in the 2.304-cell socioeconomic tables to year $t$ , using the growth rates $\rho_\sigma^c$ . Denote these updated probabilities for zone $z$ as $p_{\sigma|z}$ .

(5) For each of the $N$ households, sample a cell $\sigma$ from the 2.304-cell socioeconomic table with probabilities $r_{\sigma|z}$ . The sampling weight is $p_{\sigma|z}/r_{\sigma|z}$ .

(6) For origin zone $z$ , compute the probabilities $p_{y|\delta z}$ for every destination zone $y$ , conditional on destination district $\delta$ and origin zone $z$ . These probabilities are obtained from the UTPP tables of zone-to-zone work trip counts, by use of equation (9).

(7) For a case in origin zone $z$ and socioeconomic cell $\sigma$, compute the destination district probabilities $p_{\delta|z\sigma}$ obtained from the linear probability model of Stage II. First the dummy variables $v_{\sigma}^{(k)}$ are generated; then the estimated coefficients $b_{\delta d}^{(k)}$ are used to compute the $p_{\delta|z\sigma}$ via equations (6) and (8) where $d$ is the origin district containing zone $z$). Note that if $\sigma$ corresponds to a household with no workers, steps (7)--(9) are skipped.

(8) For this case in origin zone $z$ and socioeconomic cell $\sigma$, sample a destination district $\delta$ with probabilities $r_{\delta|z\sigma}$. The sampling weight is $p_{\delta|z\sigma}/r_{\delta|z\sigma}$.

(9) For the sampled destination district in step (8), sample a destination zone $y$ with probabilities $r_{y|\delta z}$. The sampling weight is $p_{y|\delta z}/r_{y|\delta z}$.

Next, return to step (7) and repeat for all cases in origin zone $z$. Then return to step (6), and repeat for all origin zones.

(10) For each case in socioeconomic cell $\sigma$, select randomly (with equal probabilities) one of the PUS cases in this cell. Each PUS case record consists of a housing line, followed by a person line for each person in the housing unit.

(11) For each PUS case selected in step (10), select randomly (with equal probabilities) one of the workers,[1] if any, in the

[1] A worker is identified from the corresponding PUS person line as being currently in civilian employment.

household. This step is repeated if the PUS case has been selected for more than one synthesized case.

(12) If the selected PUS household contains $m$ workers, then the sampling weight for work trips is (for $m > 0$)

$$(11) \qquad W_{yz\sigma} \equiv \frac{p_{y|\delta z}}{r_{y|\delta z}} \cdot \frac{p_{\delta|z\sigma}}{r_{\delta|z\sigma}} \cdot \frac{p_{\sigma|z}}{r_{\sigma|z}} \cdot \frac{\pi_{z|d}}{r_{z|d}} \cdot \frac{\pi_{d}}{r_{d}} \cdot m$$

and the sampling weight for non-work trips is

$$(12) \qquad N_{z\sigma} \equiv \frac{p_{\sigma|z}}{r_{\sigma|z}} \cdot \frac{\pi_{z|d}}{r_{z|d}} \cdot \frac{\pi_{d}}{r_{d}} \quad ,$$

where it is assumed that the work trip model is for individuals and the non-work trip model is for households. An output record is then assembled, containing the origin zone $z$, the destination zone $y$ (if the household contains a worker), the socioeconomic cell $\sigma$, the work and non-work trip sampling weights, a pointer to the selected worker, and the entire record for the PUS case. This synthesized case record is added to the output file. Steps (10)--(12) are repeated until all $N$ cases have been generated.

In practice, the maximum number of cases $N$ that can be generated by one pass through this procedure may be limited by the amount of computer memory available. If necessary, repeat the entire procedure (starting with the previous final setting of the random number generator) until the desired number of cases has been obtained.

If in each stage the sampling probabilities are taken to equal the universe probabilities, a random representative sample results, with weights $W_{yz\sigma} = m$ for work trips and $N_{z\sigma} = 1$ for non-work trips. Alternately, where the impacts of a policy are primarily local, the user may wish to concentrate the synthesized sample in the major impact area.

Note that PUS cases are matched with synthesized cases only on the basis of socioeconomic characteristics represented by $\sigma$ ; in general, the county of residence and the workplace district coded in the PUS record will not correspond with the selected origin and destination zones. As indicated earlier, step 10 is justified if the distribution of the socioeconomic characteristics conditioned on $\sigma$ is independent of residence and work location. This may not be the case for certain "person" variables, such as mode of transportation to work, or number of hours worked per week; such variables should not be used in conjunction with zone of employment.

This sample synthesis method has been used to generate a representative sample of 12,000 households for the San Francisco Bay Area for the year 1976. The format of the data set is described in Appendix E.

## 7. Stage IV: Updating Income Variables

After the synthesized sample file is constructed, variables such as wage rates and income which will tend to grow at uniform rates for specific occupations and industries in the Bay Area can be projected to the forecast date. We have done this by carrying out a separate econometric study of the growth rate of wages and income in the Bay Area as a function of household and wage earner characteristics.

In outline, the procedure is as follows, starting from a file of household case records (in the present case, the 1970 Census PUS file). A log-linear regression is used to estimate family income in 1970 as a function of variables which include: average earnings in the industry of each worker in 1970; average earnings in the occupation of each worker in 1970; and median income in the county of residence in 1970. Average earnings data may be obtained from state and local compilations, for example the Bay Area Salary Survey which annually gives hourly earnings in key occupations in the San Francisco Bay Area. Average hourly earnings figures are converted to annual incomes, using the subject's reported hours and number of weeks worked.

Growth rates for these variables are then estimated separately (or computed, if the relevant statistics for year t are already available). Using the estimated coefficients in the income regression, one then obtains estimates of the growth rates for income by industry, occupation code. and country of residence.

Additional points to be taken into account include variation over time in the number of workers per household and in the number of hours

worked per year, particularly in the case of workers who are not heads of household.

This stage of the synthesis has not yet been completed, and details of the method will be reported elsewhere.

## Appendix A. The Iterative Proportional Fitting Method

### A.1 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: $p_{i+} = \Sigma_j p_{ij}$ and $p_{+j} = \Sigma_i p_{ij}$. Generalizing this notation, assume $K$ variables, indexed by the elements of $\mathcal{K} = \{1,\ldots,K\}$. Let $I_k$ denote the set of categories defined for variable $k$. A cell in the contingency table is indexed by a vector $\sigma = (i_1,\ldots,i_k)$; the set of possible cell indices is denoted $S = I_1 \times \ldots \times I_k$. Cell probabilities are denoted by $p_\sigma$. 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) $Q \subseteq \mathcal{K}$; i.e., the probability of sets of the form

$$B = \bigtimes_{\ell \in Q} \{i_\ell\} \times \bigtimes_{\ell \notin Q} I_\ell \quad . \tag{A.1}$$

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

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

$$\log p_{\sigma} = \sum_{Q \subseteq K} u_{Q(\sigma)} \quad , \tag{A.2}$$

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

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

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

$$\sum_{\sigma_k \in I_k} u_{Q(\sigma)} = 0 \quad , \tag{A.4}$$

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.

## A.2 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 $Q_1,\ldots,Q_J$ . We assume each of these configurations to be maximal in the sense that none are contained in another. For each $\sigma \in S$ , let $\sigma_j$ denote the index vector formed by replacing the components $k$ of $\sigma$ for which $k \notin Q_j$ by "+" . Then, $\sigma_j$ indexes the cell of the marginal distribution with configuration $Q_j$ which is the projection of $\sigma$ . 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

$$\hat{p}_\sigma^{(i+1)} = \hat{p}_\sigma^{(i)} \frac{\bar{p}_{\sigma_j}}{\hat{p}_{\sigma_j}^{(i)}} , \quad (\sigma \in S) \tag{A.5}$$

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

---

*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 $Q_3 = \{1,2\}$.

$\hat{p}_{\sigma}^{(0)}$ have the same effects for orders exceeding the highest order marginal.

One difficulty which arises in implementation of the method outlined in Section 2 is that a full description of socioeconomic characteristics, residence, and work location are likely to yield a table with an extremely large number of cells. It is then necessary to modify the procedure to simplify or reduce the data manipulation required. The following methods are often useful:

1. Relabeling — when two variables always appear in pairs in each observed maximal configuration, they can be redefined as a single variable (with a number of categories equal to the product of the number of categories of each), reducing the dimensionality of the table, but not the total number of cells.

2. Partitioning — when a variable appears in every maximal configuration, then application of the iterative proportional fitting technique to each sub-table conditioned on a particular category of the common variable yields the same result as fitting the complete table. Hence, fitting of the full table can be partitioned into fitting of the conditional sub-tables.

3. Fitting by Stages — when a variable appears in only one maximal configuration, then the full table can be partitioned by categories of this variable, and iterative proportional fitting applied to the conditional sub-tables.

4. Stand-alone Variables — when a variable appears alone (i.e.; $\{k\}$ is a maximal configuration for variable k ), and the initial trial table $\hat{p}_{\sigma}^{(0)}$ contains no effects beyond first order involving this variable, then iterative proportional fitting can be applied to the marginal table formed by summing over the variable in question, with the final fitted probabilities for the full table given by the product of the marginal probabilities from this table and the marginal probabilities for the stand-alone variable.

## Appendix B. Comparison of the BATSC Sample with Census Counts

We have indicated in the brief description in Section 3.2 that the BATSC sample does not appear to be representative. For example, comparison of marginals for ten census tracts in San Francisco reveals substantial discrepancies as shown in Table 6.

The comparison of income statistics is obscured because of the difficulty of commensurate categorization of real incomes. However, the remaining marginals suggest a strong bias in BATSC data away from primary individuals, away from households without automobiles, and away from households with one worker. This suggests that the BATSC data is strongly biased toward higher income households which tend to have a family member at home.

TABLE 6. Comparison of BATSC with Census Data

(Selected marginal totals for ten census tracts in San Francisco)

| | 1960 Census | 1970 Census | 1965 BATSC |
|---|---|---|---|
| Family type | | | |
| Husband and wife, under 45 | 19 | 18 | 26 |
| Husband and wife, over 45 | 20 | 19 | 52 |
| Other | 61 | 63 | 22 |
| | 100% | 100% | 100% |
| Number of autos | | | |
| 0 | 56 | 54 | 30 |
| 1 | 37 | 38 | 53 |
| 2+ | 7 | 8 | 17 |
| | 100% | 100% | 100% |
| Number of workers | | | |
| 0 | 5 | 4 | 17 |
| 1 | 82 | 81 | 52 |
| 2+ | 13 | 15 | 31 |
| | 100% | 100% | 100% |
| Income* | | | |
| Lower | 61 | 47 | 34 |
| Middle | 22 | 17 | 28 |
| Upper | 17 | 36 | 38 |
| | 100% | 100% | 100% |

*The middle income bracket was $5,000 — $8,000 in 1960, $8,000 — $10,000 in 1970, and $7,000 — $10,000 for BATSC.

## Appendix C. Notes on the IPF Program

### C.1 Introduction*

This program of iterative proportional fit of the marginal totals of a contingency table, adapted from an algorithm due to Haberman (1974), is appropriate especially for the classical purpose of fitting a contingency table by combining the information from two or more sets of data.

The following are a few additions to the algorithm of Haberman:

a. An option of fitting multiple sets of marginals to a single (initial) table is available.

b. A simple consistency checking of marginal tables (input) is provided.

c. In addition to printing of some important results, two output options are available. One is for printing two-way contingency tables of the first two variables and the other for punching out the results.

Although the present program is quite flexible for the purpose of SYNSAM, it is expected that some changes should be made for other specific purposes.

---

*See Appendix A and Haberman (1974) for the general description of the algorithm.

## C.2 Input Data

The following are descriptions of input cards.

TABLE 7. Input Cards for IPF

| List of Variables in READ Statement | Format | Description |
|---|---|---|
| NVAR,NCON,NMAR,NU,NTAB, NSET,IPAIR | 7I3 | NVAR = Number of variables in the full table<br>NCON = Number of marginal tables to be fit<br>NMAR = Total number of cells in marginal tables (= dimension of IMARG)<br>NU = Dimension of work area U used to store fitted marginal totals<br>NTAB = Number of cells in the full table<br>NSET = Number of sets of fitted tables to be obtained<br>IPAIR= 2 if fitted tables at two different dates per set are to be obtained<br>= 1 if only one fitted table per set is to be obtained |
| POPT2,POPT3 | 2I2 | POPT2= 1 if two-way contingency tables are to be printed<br>= 0 otherwise<br>POPT3= 1 if fitted relative frequencies are to be punched out<br>= 0 otherwise |
| (CONFIG(J,IP),J=1,NVAR) (Repeated for IP=1,NCON) | 10I2 | The sets indicating marginal totals to be fit.* |
| (DIM(I),I=1,NVAR) | 10I2 | The number of categories of each variable in the full table. |

TABLE 7 continued

| List of Variables in READ Statement | Format | Description |
|---|---|---|
| (LOCMAR(I),I=1,NCON) | 20I3 | Pointers to the table in IMARG* |
| (FIT(I),I=1,NTAB) | 8F10.0 | Initial table**,* |
| (VLAB(I),I=1,NVAR) | 10A7 | Labels of variables; input if POPT2=1 |
| (CLAB(I),I=1,N1) | 10A7 | Labels of categories of each variable with N1 = ΣDIM(I) ; input if POPT2=1 |
| (IMARG(I),I=1,NMAR) | 8I10 | Marginal tables to be fit*,** |

*See Haberman (1974) for its construction.

**This input could be handled through tape if NTAB or NSET is large. The output also may be stored on tape.

IPAIR may be set to 2 especially when two fitted tables of different dates are to be printed out in the form of two-way contingency tables for easy comparison. Other cases can be handled by setting IPAIR = 1 and increasing NSET as desired. For example, when census marginals of five counties in 1960 and 1970 are to be fit by using 1970 PUS full table as the initial, we may put NSET = 5 and IPAIR = 2 . Thus 10 sets of marginals are supposed to be read in with the following ordering:

$$(IMARG_1)^{1960}, (IMARG_1)^{1970}, \ldots, (IMARG_5)^{1960}, (IMARG_5)^{1970}$$

Standard use of iterative proportional fit, as opposed to the classical use, may be accomplished when every element of initial table (FIT) is set to 1 (assuming complete table). (Haberman (1974)).

## C.3 Program Structure

The program has a simple structure which calls 6 subroutines including 2 important iterative computational routines, COLLAP and ADJUST . The following are brief descriptions of the subroutines.

### SUBROUTINE PRIOR

This routine checks the inputs, NVAR, NTAB, DIM, NU, NMAR, and IMARG for their validity and mutual consistency before doing any data manipulation. Following error indicators are used:

Error number 1 ; CONFIG or LOCMAR contains errors,

Error number 2 ; NTAB, DIM, NU, or NMAR contains errors,

Error number 4 ; NVAR is not valid.

### SUBROUTINE MARCHEK

This is the routine for checking the consistency of marginals (input). Error number 3 is printed and the program stops when there is inconsistency.

### SUBROUTINE COLLAP

This subroutine computes the marginal tables of each configuration of the previous fitted table or the initial table for the first iteration. Thus it computes $\hat{P}_{\sigma_j}^{(i)}$ of equation (A.5) where $\sigma_j$ covers one configuration.

SUBROUTINE ADJUST

This makes proportional adjustment according to equation (A.5) with the result of SUBROUTINE COLLAP. For each set of marginals (IMARG) these two subroutines are repeated either until the maximum cell deviation observed between two successive iterations becomes less than MAXDEV which is set to 0.5 in the program, or until MAXIT, which is the maximum number of iterations set in the program by data declaration, equals 20. Thus, depending upon the size of NTAB and total number of cell counts, MAXDEV and MAXIT could be adjusted.

SUBROUTINE MAR

This subroutine prints out the marginals in the order of configuration.

SUBROUTINE MAT

This is to print the fitted frequencies and relative frequencies in the form of two-way contingency tables of the first two variables. In the present program the maximum number of categories of the second variable for this routine to work is three. With changed formats other cases can be handled. If IPAIR=1 , it prints tables of fitted result and initial input. If IPAIR=2 , then it prints tables of two fitted results without the initial input. Since the number of two-way tables is NTAB/DIM(1)*DIM(2) , the option POPT2 must be used with some caution.

# Appendix D. Notes on the Sampling Program

## D.1 General Features of the Program

### D.1.1 Implementation

The sampling procedure has been implemented as a FORTRAN program and subroutines. It was designed for use on the CDC 7600 computer at Lawrence Berkeley Laboratory (LBL) to generate synthetic samples of 12,000 (or fewer) cases for the San Francisco Bay Area. Although the program has been made flexible enough for more general application, a few changes will probably be required before it is able to generate samples for other areas, utilize alternative data sources, or run under different operating systems.

Application to another SMSA or County Group may involve changing the dimensions of certain arrays in the program (see section D.4.3 below). Use of the program at another computer installation will require name changes in calls made by the program to certain system library routines, and possibly also the elimination of non-standard FORTRAN features such as end-of-file checking (see section D.4.4 below). New versions of the subroutines that read and write Public Use Sample case records will also be required, since these subroutines are designed for the special internal block structure and re-coded binary tabulations of the census tapes available at LBL (see section D.3).

D.1.2 Sampling Weight Options

The program provides options for specifying the sampling probabilities for origin districts ($r_d$) and for origin zones ($r_{z|d}$) in steps (2) and (3) of the sampling procedure described in section 6.2. At present no options are available for user-specified sampling probabilities for sampling on cells, destination zones, or destination districts: in these cases the sampling probabilities are set equal to the universe probabilities (i.e., $r_{...} = p_{...}$) .

The options for origin sampling probabilities, which may be specified independently for districts and for zones, are as follows:

Option 0: no sampling probabilities are provided. The projected distribution of households is used to obtain the probabilities.

Option 1: a set of weight factors is provided as an input. The sampling probability is the universe probability ($p_d$ or $p_{z|d}$) multiplied by the weight factor for that district or zone.

Option 2: the sampling probabilities are supplied by the user as an input.

D.1.3 Definition of Socioeconomic Cells

A control card in the input deck allows one to change the number of socioeconomic characteristics used to define the cells. In the present

application, for example, this card reads

9 3 4 3 2 2 2 2 2 2

where 9 is the number of characteristics used, followed by a list of the number of alternatives that each of these nine characteristics can take (cf. Table 2). Obviously the number of cells (2304) is the product of these nine numbers, while the number of dummy variables (13) is obtained by subtracting 1 from each of the nine numbers and then adding. Including a coefficient for the constant term then gives a total of 14 coefficients in the linear probability model for destination districts.

It is assumed that cells are enumerated and dummy variables are defined according to the following rules:

(1) Cells are enumerated by varying the first characteristic first.

(2) The first characteristic is always "number of workers in the household," and the first alternative for this characteristic is "zero workers." Thus, in the present application, cells 1, 4, 7, ..., 2302 correspond to households with zero workers, and no destination will be assigned to cases in these cells.

(3) In forming dummy variables, the last alternative for each characteristic is dropped.

(4) The first regression coefficient in the linear probability model corresponds to the constant term, followed by the dummy variables for characteristic 1 , etc.

D.1.4 Treatment of Missing Data and Empty Cells

The sampling procedure outlined in Section 6.2 has to be slightly modified to guard against selecting a zone, district or cell for which the conditional probabilities required in a subsequent step cannot be calculated. For example, it can happen that no work trips are recorded on the UTPP tape from an origin zone $z$ to any of the zones $y$ in some destination district $\delta$, and thus the conditional probabilities $P_{y|\delta z}$ are undefined, even though the probabilities for selecting $\delta$ and $z$ may be nonzero.

The following modifications are therefore made to the sampling probabilities defined in Section 6.2.

In step (1):

(a) If the socioeconomic cell probabilities for zone $z$ could not be computed by the IPF program in Stage I, then set $\pi_z$ to zero instead of $n_z/n$. A list of such zones is supplied as an input to the program. (It is assumed that the records corresponding to these zones are physically present on the input tape, but do not contain meaningful data.)

(b) If there are no work trips originating in zone $z$ recorded on the UTPP tape, then set $\pi_z$ to zero instead of $n_z/n$. A list of such zones is supplied as an input to the program. (It is assumed that the records corresponding to these zones were omitted from the input tape.)

(c) The remaining $\pi_z$ are rescaled so that they sum to one.

(d) The $\pi_d$ are obtained by summing these "adjusted" values of $\pi_z$.

In step (2):

(a) If option 1 or 2 has been chosen for the origin district sampling probabilities, denote by $w_d$ the specified weights or probabilities. Let $w_z$ be the analogous quantities for origin zones. Before sampling on origin districts, we first compute un-normalized zone "probabilities" $r_z$, defined as follows according to the zone sampling option.

Option 0: $r_z = \pi_z$

Option 1: $r_z = \pi_z \cdot w_z$

Option 2: $r_z = \begin{cases} w_z & \text{if } \pi_z \neq 0 \\ 0 & \text{if } \pi_z = 0 \end{cases}$ .

(b) District "probabilities" $r_d$ are defined similarly, according to the district sampling option. However, if $r_z$ is zero for all zones in an origin district $d$, then $r_d$ is set to zero.

(c) The $r_d$ are rescaled so that they sum to one.

In step (3), the origin zone sampling probabilities are computed in terms of the $r_z$ defined above:

$$r_{z|d} = r_z / \sum_{z' \in D_d} r_{z'} \quad ,$$

but are not calculated for any district with $r_d = 0$.

In step (4), the previously-constructed index to the PUS tape is read in before updating the socioeconomic contingency tables. This index gives the number of PUS cases in each socioeconomic cell. If there are no PUS cases in cell $\sigma$, then for every zone $z$, $p_{\sigma|z}$ is set to zero. The remaining probabilities are rescaled to sum to one.

In step (6), the conditional destination zone probabilities $p_{y|\delta z}$ are not computed if there are no trips recorded on the UTPP tape from zone $z$ to destination district $\delta$. In Equation (9) this corresponds to $p_{y|z} = 0$ for all $y \in D_\delta$.

In step (7), the estimate $p'_{\delta|z\sigma}$ obtained from Equation (6) is replaced by zero if the conditional zone probabilities $p_{y|\delta z}$ could not be computed in the previous step. (If the resulting $p'_{\delta|z\sigma}$ are negative for every $\delta$, then equation (8) is replaced by the following procedure. Of those destination districts $\delta$ for which $p_{y|\delta z}$ could be calculated in step (6), select the one with the maximum value of $p'_{\delta|z\sigma}$. For this district, $p_{\delta|z\sigma}$ is set equal to one.)

In step (10), random selection of a PUS case occurs only if there is more than one PUS case in the cell, and in step (11) random selection of a worker occurs only if there is more than one worker in the household.

In step (12), the sampling weight for work trips reduces to

$$W_z = \frac{\pi_z}{r_{z|d} \cdot r_d} \cdot m$$

because there are user-defined sampling probabilities only for origin districts and zones. A zone sampling weight factor

$$S_z = \frac{\pi_z}{r_{z|d} \cdot r_d}$$

is therefore calculated in step (3) for all zones for which $r_{z|d}$ and $r_d$ are nonzero.

## D.2 Input Data for the Sampling Program

### D.2.1 Input Tape

Input data for the program is on cards and on binary (i.e., unformatted) tape. All the data on input tapes shown in Figure 4 (IPF tables and growth factors; UTPP tables; sorted PUS case file) was consolidated on a single input tape.*

The figures given here apply to the SYNSAM tape for the San Francisco Bay Area, with 440 zones, 6 origin districts, 13 destination districts, 5 "counties," and 2304 socioeconomic cells. These parameters can be changed by control cards in the input deck (see Section D.2.2 below). The input tape contains:

(1) The PUS index table, giving the number of PUS cases in each socioeconomic cell: 1 record of 2304 words (INTEGER).

---

*The input tape used in the present application contains approximately 1.8 million words.

(2) The IPF growth factors, giving growth rates for each of 2304 socioeconomic cell probabilities for each of 5 "counties" (as defined in Section 4): 5 records, each of 2304 words (REAL). Note that in fact the inputs were not the annual growth factors $\rho_\sigma^c$, but rather the five-year growth factors

$$\rho_\sigma^c(5) = \sqrt{\frac{p_\sigma^c(1970)}{p_\sigma^c(1960)}} \quad .$$

(3) The IPF fitted tables, giving 2304 socioeconomic cell probabilities for each of 440 zones: 440 records, each of 2304 words (REAL). Note that in fact the inputs were probabilities projected to 1975, i.e., $p_\sigma^c(1975)$, rather than the base-year probabilities $p_\sigma^c(1970)$. The required probabilities are therefore computed in the program as

$$p_\sigma^z(t) = A_z(t) \cdot p_\sigma^z(1975) \cdot [\rho_\sigma^c(5)]^{(t-1975)/5}$$

where the $A_z(t)$ are such that the probabilities for each zone sum to one.

(4) UTPP tables, giving the number of trips to each of 440 destination zones from each of 440 origin zones:* 437 records, each of 440 words (INTEGER). These tables were compiled by extracting and consolidating counts from the 1970 Census UTPP file.

(5) PUS case records, previously sorted by socioeconomic cell

---

*Three origin zones were empty.

number*: 1029 records, each of 504 words (INTEGER). The original PUS tapes have been recoded at LBL (see Section D.3 below), the unit records being converted to a binary code and then assembled into 504-word blocks. This recode has been retained in the sampling procedure, which copies PUS case records to the output tape without alteration. However, alternative versions of the input and output subroutines are available which will handle case records in the original coded format of the Census Bureau tapes (1970 PUS, SMSA 15% sample). The 1970 PUS for the San Francisco-Oakland and San Jose SMSA's contains approximately 14,000 households.

The file structure of the input tape is as follows. Although items (1) - ( 5) above have been consolidated on the same tape, each is a separate logical file and is terminated by a file mark. At the end of each of items (1) - (4), the program checks that an end-of-file condition is present and, if so, clears it (i.e., prepares to read the next file). If an end-of-file does not occur where expected, or occurs in an unexpected place, an error condition is raised and the program will terminate. In addition, the set of 440 fitted socioeconomic tables was divided into 15 separate logical files, for technical reasons. This number is specified on an input card (see Section D.2.2 below) and is stored in variable MFILE; the program checks that the correct number (MFILE-1) of end-of-file conditions occurs while reading the socioeconomic tables.** If the set of socioeconomic tables is not divided into files, MFILE is set equal to 1 . Thus the input tape contains MFILE + 4 logical files.

---

*Vacant units and individuals in group quarters were removed.

**The actual number of tables in each of these files is arbitrary and need not be specified; only the total numbers of files and tables are checked.

The syntax for end-of-file checking is compiler-dependent. We note also that an operating system might not allow a FORTRAN program to skip past an end-of-file and continue reading the next file on the same input unit. Thus, if the program is to be run elsewhere, one of the following modifications may be found preferable:

(a) Consolidate input data items (1) - (5) on the tape as a single logical file or data set. The end-of-file checks in the program can then be deleted.

(b) Take each input data item (1) -(5) as a separate single logical file or data set, and assign each to a separate input unit for the FORTRAN program. The READ(1) statements in the program are modified accordingly, and the end-of-file checks deleted.

D.2.2 Input Cards

The input deck contains: control cards; definitions of origin districts, destination districts and counties in terms of zones; numbers of households and growth rates by zone; lists of empty and missing zones; sampling weight options and user-supplied weights (if required) for origin districts and origin zones; and the estimated coefficients of the linear probability equation for destination districts. Details and formats* are given in Table 8.

---

*In F-format specifications, the number of decimal places is arbitrary because it can be overridden by the position of a decimal point in the input field.

TABLE 8. Input Cards for the Sampling Program

| List of Variables in READ Statement | Format | Description of Variables |
|---|---|---|
| (1) Defining parameters | | |
| NCASES,ISEQ | (2I5) | Number of cases to be generated; sequential identifying number of the first case. |
| DEBUG,TEST | (7L1) | Six debugging options and an option for listing output cases, set "on" by T and set "off" by F (see Section D.5). |
| (IW(I),I=1,6) | (6I12) | Six initial seed values supplied for the random number generating routine RGEN . If IW(6) is zero, this input is ignored and default seed values are automatically supplied. |
| YEAR | (F4.0) | Year (for updating PLUM and IPF data) |
| NZONES,NODIST,NDDIST, NCNTY,MFILE | (5I5) | Number of zones; number of origin districts; number of destination districts (excluding "sink"); number of "counties" (i.e., regions used in defining IPF cell growth factors); number of files into which the IPF fitted socioeconomic tables have been divided (see Section D.2.1) |
| NCVARS,(NALT(I),I=1, NCVARS) | (16I5) | Number of socioeconomic characteristics used to define cells; number of alternatives that each of these characteristics can take (see Section D.1.3). |

TABLE 8 continued

| List of Variables in READ Statement | Format | Description of Variables |
|---|---|---|
| (2) District and "county" definitions | | |
| N,(IW(J),J=1,N) (repeated for each origin district) | (16I5) | For each origin district: number of zones in this district; list of zones in this district, not necessarily in numerical order. |
| N,(IW(J),J=1,N) (repeated for each destination district) | (16I5) | For each destination district: number of zones in this district; list of zones in this district, not necessarily in numerical order. |
| N,(IW(J),J=1,N) (repeated for each "county") | (16I5) | For each "county": number of zones in this "county"; list of zones in this "county," not necessarily in numerical order. |
| (3) Zone populations and growth percentages | | |
| (ZONEWT(I),I=1,NZONES) | (16F5.0) | Numbers of households in each zone, 1970. |
| (W(I),I=1,NZONES) | (16F5.2) | Projected 10-year growth percentages for numbers of households in each zone, 1970-1980. |
| (4) Empty and missing zones | | |
| N,(IW(J),J=1,N) | (16I5) | Number of empty zones on IPF tape; list of such empty zones, not necessarily in numerical order |

TABLE 8 continued

| List of Variables in READ Statement | Format | Description of Variables |
|---|---|---|
| N,(IW(J),J=1,N) | (16I5) | Number of missing zones in UTPP tables; list of such missing zones (must be in numerical order). |
| (5) Origin sampling weights (if required) | | |
| IDWT | (I5) | Origin district sample weight option (Section D.1.2). |
| (DISTWT(I),I=1,NODIST)<br>(not present if IDWT.EQ.0) | (16F5.0) | Origin district sample weights, if present. |
| IZWT | (I5) | Origin zone sample weight option (Section D.1.2). |
| (W(I),I=1,NZONES)<br>(not present if IZWT.EQ.0) | (16F5.0) | Origin zone sample weights, if present. |
| (6) Regression coefficients | | |
| (COEFF(I,J,K),I=1,NV)<br>(repeated for each origin district-destination district pair) | (8F10.0) | Estimated coefficients for the linear probability model for destination districts. Primary ordering is by origin district K; within each origin district the ordering is by destination district J . Coefficients start with the coefficient of the constant term (I=1). See also Section D.1.3. |
| (7) Terminator | | |
| A | (A1) | Last card contains $ in column 1 (to check that correct number of cards has been read). |

Note that the growth rates for numbers of households were actually supplied as percentage increases projected for each zone from 1970 to 1980, say $g_z$ . The required numbers are therefore computed in the program as

$$n_z(t) = n_z(1970) \cdot (1 + g_z/100)^{(t-1970)/10} .$$

## D.3 Binary Coding of 1970 Census PUS Case Records

Census tapes have been recoded at LBL in binary form. The recoding and block structure of these tapes have been written up on the LBL data cell, library CENSUS, subset WRITEUP (as of October 1976). Tables are given there which give the bit positions in the recode corresponding to each census tabulation.

This binary recode was retained for the PUS case records handled by the sampling program. This main features are as follows.

(1) Each PUS case consists of a household line, followed by a person line for each person in the housing unit. On the original census tapes each line is coded (formatted) with 120 character positions; the number of person lines to follow is given in positions 12 and 13 of the housing line ("tab H12-13"). In the binary recode, each line becomes 10 words (of 60 bits each); the second word in the household line is an integer giving the number of person lines.

(2) As many cases as will fit are grouped in a 500-word "block,"

which thus contains up to 50 census lines. Any remaining lines at the end of the block are unset.

(3) Each block is preceded by a four-word directory, of which the second word gives the number of census lines in the block. The directory and block together form a 504-word logical record on the tape.

This applies only to programs run at LBL. Subroutines for reading and writing conventionally formatted PUS records are available for use elsewhere, if required.

## D.4 Program Structure

### D.4.1 Description of Principal Routines

The subroutine structure of the sampling program is shown in Figure 6. (Calls to certain system library routines are not shown, but are listed in Section D.4.4 below.) The routines are as follows.

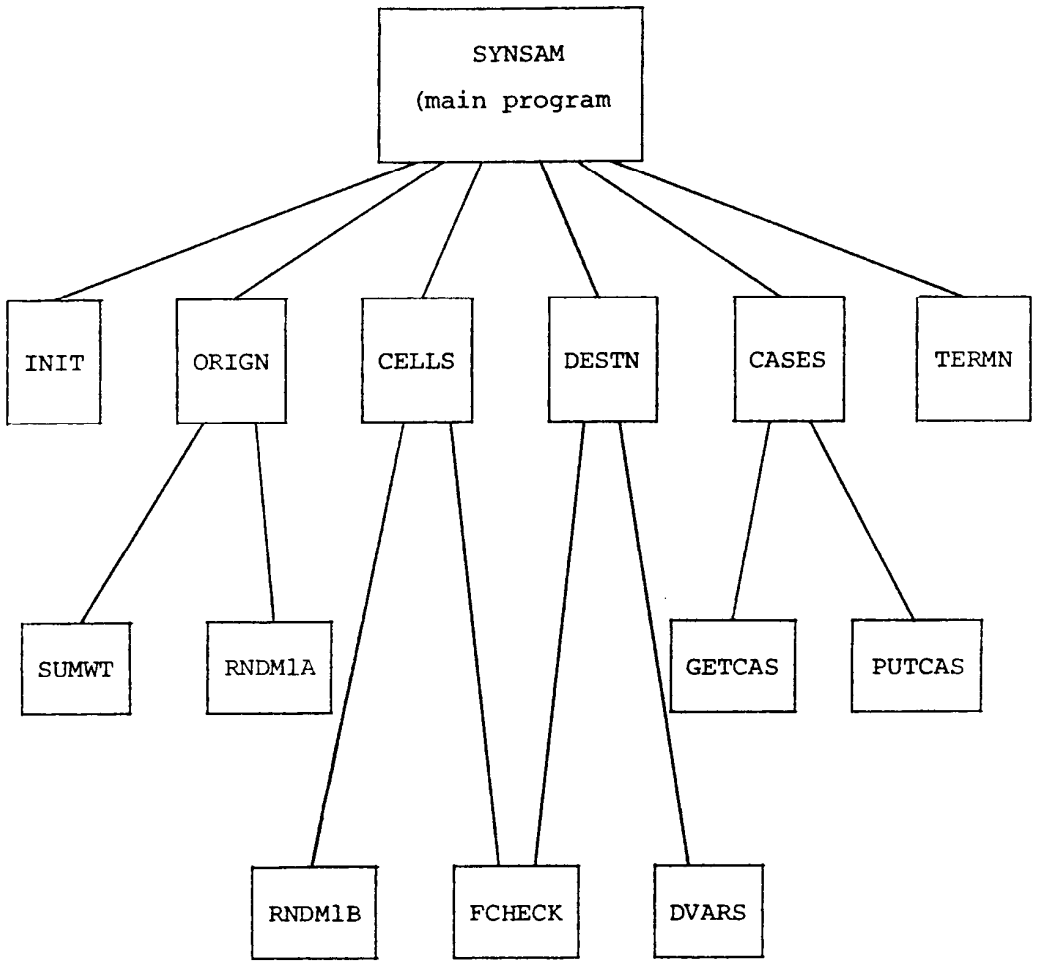
PROGRAM SYNSAM*

The main program is essentially a holding routine, in which array dimensions are set. The only executable statements are calls to various subroutines. The contents and dimensions of the principal arrays are

---

*The PROGRAM statement is used only on CDC machines.

FIGURE 6. Subroutine Structure of the Sampling Program (excluding system library calls)



given in Table 9 (see Section D.4.3).

The input/output unit numbers are as follows:

1: binary input tape* (see Section D.2.1);

2: binary output tape (see Appendix E);

5: input card deck (see Section D.2.2 and Table 8);

6: printed output messages and debugging information.

SUBROUTINE INIT

This routine reads and checks the control cards, the defining parameters, the district and county definitions in terms of zones, the zone populations and growth rates, and the lists of empty and missing zones (see Table 8). It implements step (1) of the sampling procedure.

An option is provided for reading in specified seed values for the random number generator, instead of using the standard default starting values.** This would be useful if a large sample were being generated by successive runs of the program, for example. For the same reason, the sequence number of the first case can be specified as an

---

*An SCM buffer size of 4401 (octal) was specified for TAPE1, and the program was run with LCM buffer sizes of 200K (octal) for TAPE1 and TAPE2. This applies to the CDC 7600 computer only.

**If new seed values are specified for the random number generator, it is advisable to use values obtained at the end of a previous run; it is possible that a non-random sequence might result if seed values are badly chosen.

input (the default value is 1).

The defining parameters are first checked to ensure that they are compatible with the array dimensions set in the main program. A set of index tables is set up: origin and destination districts are listed by zone, zones are listed by origin district and by destination district, and counties listed by zone. The projected number of households in each zone is calculated for the specified year, and the projected total number of households for the entire area is printed out. In the case of empty and missing zones, the populations are then reset to zero.

Out-of-range parameters and certain other inconsistencies lead to appropriate error messages and program termination. There are 18 such messages which are in general self-explanatory.

SUBROUTINE ORIGN

This distributes the cases among origin districts and then zones, as described for steps (2) and (3) of the sampling procedure. User-supplied sampling weights or probabilities are read in if the appropriate options are selected (see Table 8). The routine returns array KZONES, giving the number of cases in each origin zone, and array SZNWT, giving the sampling weight factor associated with each origin zone.

SUBROUTINE CELLS

This carries out steps (4) and (5) of the sampling procedure. The PUS index table, the IPF growth factors, and the IPF fitted tables are read in. Socioeconomic cell probabilities are updated to the required

year, and a cell is picked for each case. The routine returns array KCELLS, containing the cell number of each case.

SUBROUTINE DESTN

This carries out steps (6) - (9) of the sampling procedure. The estimated coefficients for destination districts, and the UTPP trip tables, are read in and are used to calculate destination probabilities. The conditional destination zone probabilities $p_{y|\delta z}$ are first calculated and stored in array ZONEWT for each origin zone $z$, before going through the cases in that zone. A call to subroutine DVARS provides the vector of dummy variables associated with cell $\sigma$, and the destination district probabilities $p_{\delta|z\sigma}$ are then calculated. For each case, a destination district and then a zone are picked. The routine returns array KDESTN, containing the destination zone for each case. Zero is returned for households with no workers.

SUBROUTINE CASES

This completes the sampling procedure, steps (10) - (12). The actual input and output of PUS case records are handled by calls to subroutines GETCAS and PUTCAS respectively. A search is first made for all synthesized cases in socioeconomic cell IC. For cases in this subset, temporary arrays IW, LW and JW contain respectively the case number, the origin zone, and a randomly selected integer IHH in the range $1 \leq IHH \leq NHH$, where NHH is the number of PUS cases in this cell. The

PUS cases in cell IC can then be processed in input sequence: each is attached to all the synthesized cases (if any) with the corresponding value of IHH .

Before each synthesized case is sent to subroutine PUTCAS for output, a worker in the household is randomly selected (if there are two or more workers), and the sampling weights computed. Note that the sequence of output cases is different from the sequence of synthesized cases in arrays KCELLS and KDESTN .

SUBROUTINE TERMN

This prints out some concluding messages, including the final seed values in the random number generator.

## D.4.2 Description of Subsidiary Routines

The following routines are not directly involved in the sampling procedure.

SUBROUTINE SUMWT(ZWT,DWT,INDX,NZ,ND)

This returns district weights in DWT , given a set of zone weights in ZWT . Parameter NZ is the dimension of arrays ZWT and INDX , while ND is the dimension of array DWT . Array INDX gives the district corresponding to each zone.

SUBROUTINE RNDM1A(N,M,P,Q,LIST)

This routine generates a frequency distribution of multinomial random variables. It is intended for use when N , the sample size, is larger than M , the number of alternatives. Individual random numbers are obtained by function calls to RGEN , a random number generator. The remaining parameters are:

P is the array containing the M probabilities, supplied as an input. These are assumed to be non-negative and to sum to one.

LIST is the array which returns the M components of the generated frequency distribution.

Q is a working space array, with dimension M.

This routine is used for distributing cases among origin districts and zones.

SUBROUTINE RNDM1B(N,M,P,Q,IQ,LIST)

This is essentially the same as the previous routine RNDM1A, except the resulting distribution is returned in a different form and a different sorting method is used. It is intended for use when N , the sample size, is smaller than M , the number of alternatives. The other parameters are:

P is the array containing the M probabilities, supplied as an input (assumed to be non-negative and correctly normalized).

LIST is the array which returns the selected alternative for each of the N cases in the sample (sorted by selected alternative in ascending order).

Q is a working space array of dimension N .

IQ is a working space array of dimension N+1 .

This routine is used for distributing the cases in each zone among socioeconomic cells.

SUBROUTINE DVARS(ICELL,IDV,NALT,DVINDX,NDV,NCV)

This returns the vector of dummy variables corresponding to the specified socioeconomic cell ICELL . These are defined as in Section D.1.3. Parameter NDV is the number of dummy variables (including one for the constant term), and NCV is the number of socioeconomic characteristics used in defining the cells. The remaining parameters are:

IDV is the array which returns the NDV dummy variables. Note that IDV(1) , which corresponds to the constant term, is not in fact set.

NALT is an array which specifies the number of alternatives (or categories) for each of the NCV socioeconomic variables.

DVINDX is an array of dimension NCV such that the dummy variable IDV(DVINDX(I)+J) corresponds to alternative J of socioeconomic variable I .

SUBROUTINE GETCAS(NW)

This reads the next case from the sorted PUS case file, and stores it in a COMMON block* for subsequent output by subroutine PUTCAS . Pointers are set to those person lines which correspond to workers. The number of workers in the household is returned as NW .

Alternative versions of this and the following routine PUTCAS are required, depending on whether the formatted or binary version of the PUS data is used (see Section D.3 for the binary version available at LBL). In the formatted version, each case is read in as it is required. In the binary version, records are read in blocks of up to 50 census lines; thus a read takes place only on the first call to the subroutine or if the pointer LOC passes the last non-empty line of the buffer B .

The only data read from the case records is:

(a) Number of person records (positions 12-13 in the household record; word 2 in binary form);

(b) Employment status (position 31 in the person record; bits 13-18 of word 3 in binary form), which is 1 if currently at work.

---

*In the binary input version this block is called /BUFFER/ ; in the formatted input version it is /REC/ .

There should not be more than 28 persons in any household.

SUBROUTINE PUTCAS(ISEQ,IOZ,IDZ,IC,SWT1,SWT2,IW,NW,TEST,NCASES)

For each synthesized case, this writes an output record(s) consisting of 8 variables computed by the sampling program plus the corresponding PUS case record (household and all associated person lines). The parameters are:

ISEQ: case sequence number;

IOZ: origin zone;

IDZ: destination zone (zero if NW=0);

IC: socioeconomic cell number;

SWT1: sampling weight for work trips;

SWT2: sampling weight for non-work trips:

IW: number of the selected worker (IW $\leq$ NW);

NW: number of workers in the household;

TEST: a logical variable controlling printed output. If .TRUE. , all cases are printed out; otherwise only the first and last 20 cases are printed.

NCASES: the sequence number of the last case.

Output variables calculated in the routine are: IP, the person line number corresponding to worker IW ; and NR , the number of census lines copied from PUS for this case (equal to NPERS + 1 ).

SUBROUTINE FCHECK(N)

This checks whether there is an end-of-file condition on input unit 1 . If so, the end-of-file marker is skipped; if not, a message is printed giving the value of N , and the program is terminated.

### D.4.3 Contents of Principal Variables and Arrays

Common block /PARAMS/ contains the following variables, which are read from the input deck (or which are calculated from parameters specified in the input deck) given in Table 8.

YEAR: the year for which the sample is required

NCASES: the number of cases to be generated

NCELLS: the number of socioeconomic cells

NZONES: the number of zones in the area

NODIST: the number of origin districts

NDDIST: the number of destination districts, excluding the "sink"

NCONTY: the number of counties (i.e., regions for which IPF growth factors are defined)

NCVARS: the number of socioeconomic variables

NDVARS: the number of dummy variables, including the constant

ISEQ: the case sequence number

DEBUG(6): a set of switches controlling six debugging options

TEST: a switch controlling output printing of synthesized cases

MFILE: the number of files into which the set of fitted IPF tables has been divided

NGAP: the number of missing origin zones in the UTPP tables.

The arrays in common blocks /SPACE1/ and /SPACE2/ are listed in Table 9, with a summary of their main uses. To determine what the actual dimension of one of these arrays should be, take the name of the dimension parameter and change the initial letter from M to N: this gives the name of the corresponding variable in /PARAMS/ whose value determines the minimum array size. For example, to select a suitable size for array KCELLS(MCASES) we refer to the variable NCASES: the array size must be at least equal to the number of cases to be generated.

An exception applies to the working space arrays with dimension MWSP . This dimension must not be smaller than any of:

(a) NZONES;

(b) (maximum number of cases in any one origin zone) + 1 ;

(c) (maximum number of cases in any one socioeconomic cell) + 1 .

TABLE 9. Contents of Arrays in Sampling Program

| | |
|---|---|
| CELLWT(MCELLS) | Set to 0 if the corresponding cell is empty on the PUS file; set to 1 otherwise. |
| COEFF(MDVARS,MDDIST,MODIST) | Estimated regression coefficients in the linear probability model for destination districts. |
| CYINDX(MZONES) [INTEGER] | Gives county corresponding to each zone. |
| DDINDX(MZONES) [INTEGER] | Gives destination district corresponding to each zone. |
| DDPTR(MDDIST) [INTEGER] | Points to the elements of array DZLIST containing the number of the first zone in each destination district. |
| DDSIZE(MDDIST) [INTEGER] | Number of zones in each destination district. |
| DISTWT(MDDIST) | Used for origin district and destination district sampling probabilities. |
| DSUM(MDDIST) | Temporary storage used in computing conditional probabilities for origin and destination zones. |
| DVINDX(MCVARS) [INTEGER] | Points to the locations preceding the elements of array DVARS containing the dummy variables for the first alternative of each socioeconomic characteristic. |
| DZLIST(MZONES) [INTEGER] | List of zones ordered by destination district. |
| FIT(MCELLS) | Contains the socioeconomic contingency table for the current zone. |
| IDBG4(MODIST) | Used to control output under debugging option 4. |
| IDV(MDVARS) | Dummy variables used in the linear probability model for destination districts. |
| IPU(MCELLS) | PUS index table. |

TABLE 9 continued

| | | |
|---|---|---|
| IW(MWSP) | | Working space array. |
| JW(MWSP) (*) | | Working space array. |
| KCELLS(MCASES) | | Socioeconomic cell selected for each case. |
| KDESTN(MCASES) (*) | | Destination zone selected for each case. |
| KDISTS(MODIST) | | Number of cases assigned to each origin district. |
| KW(MWSP) (*) | | Working space array. |
| KZONES(MZONES) | | Number of cases assigned to each origin zone . |
| LW(MWSP) (*) | | Working space array. |
| NALT(NCVARS) | | Number of alternatives for each socio-economic variable. |
| ODINDX(MZONES) | [INTEGER] | Gives origin district corresponding to each zone. |
| ODPTR(MODIST) | [INTEGER] | Points to the elements of array OZLIST containing the number of the first zone in each origin district. |
| ODSIZE(MODIST) | [INTEGER] | Number of zones in each origin district. |
| OZLIST(MZONES) | [INTEGER] | List of zones ordered by origin district. |
| PFGROW(MCELLS,MCNTY) (*) | | Growth factors for the socioeconomic tables. |
| SZNWT(MZONES) | | Sampling weight factor for each origin zone. |
| U(MWSP) (*) | | Working space array. |
| UTPGAP(MGAP) | [INTEGER] | List of missing origin zones in the UTPP tables. |
| V(MWSP) (*) | | Working space array. |

TABLE 9 continued

| | |
|---|---|
| W(MWSP) (*) | Working space array; also used for origin zone sampling probabilities (unnormalized). |
| WDSUM(MODIST) | Used in calculating origin zone sampling probabilities. |
| ZNPTR(MZONES) | Points to the elements of arrays KCELLS and KDESTN containing the first case in each zone. |
| ZONEWT(MZONES) | Used for origin and destination zone probabilities. |

(*) Arrays involved in EQUIVALENCEs.

Estimates of (b) and (c) for a sample of given size can be made on the basis of the household distributions by zone and the PUS index table by socioeconomic cell, with a reasonable margin allowed for statistical fluctuations. A message will be printed if the working space is exhausted.

Note that variables such as MCASES are set by DATA statements in the main program and are equal to actual array dimensions, whereas the corresponding variable NCASES is read from input data for generating a particular sample. If NCASES exceeds MCASES , an error message will be issued; but if MCASES is not set equal to the actual array dimensions then the program may give anomalous results.

D.4.4 Non-standard Features

Calls are made to the following system library routines. (If the program is run elsewhere, the calls to the random number generator will have to be changed.)

FUNCTION RGEN(I)

This returns a random number in the range (0,1) . The argument is a dummy and is ignored.

SUBROUTINE STOGEN(ISV)

This returns the current seed values in the random number generator RGEN . ISV is an array of dimension 6 , containing five seed values and an index value which ranges from 1 to 5 .

SUBROUTINE LODGEN(ISV)

This sets new seed values in the random number generator RGEN . The contents of array ISV are as above.

FUNCTION EOF(N)

If the last READ statement on input unit N did not encounter an end-of-file, this returns zero; otherwise, the end-of-file is cleared and a nonzero value is returned. This is a FORTRAN library function on CDC machines only. Section D.2.1 describes the file structure of the input tape, and ways of eliminating the need for end-of-file checking.

In addition, the binary-input version of subroutine GETCAS contains some non-standard FORTRAN coding, but this will be acceptable to machines that can read the binary form of the PUS tape.

## D.5 Debugging Options

There are 7 options for printing out various arrays during the course of the program. Each option is "on" when the corresponding logical variable is TRUE . Options are set by a card in the input deck (see Table 8). The relevant subroutines and the nature of the information provided by each option are as follows:

DEBUG(1): Subroutine INIT: projected household numbers; index tables and pointers for origin and destination districts.

Subroutine ORIGN: arrays used in computing origin district and zone sampling probabilities.

DEBUG(2): Subroutine ORIGN: number of cases assigned to each origin district and zone; origin zone sampling weight factors.

DEBUG(3): Subroutine CELLS: cell assigned to each case.

DEBUG(4): Subroutine DESTN: destination district probabilities for the first five cases in each origin district.

DEBUG(5): Subroutine DESTN: conditional destination zone probabilities for the first five origin zones selected.

DEBUG(6): Subroutine DESTN: destination zone assigned to each case.

TEST: Subroutine PUTCAS: print all synthesized case records (suppressing person records except for the selected worker in each household); otherwise only the first and last 20 cases are printed.

## Appendix E. Format of the Output Tapes

A tape has been written containing 12,000 household case records which form a representative random sample of the San Francisco Bay Area. Each case consists of 8 variables computed by the SYNSAM program, followed by data copied from the PUS tape. The PUS data for each case consists of a housing line followed by NP person lines. Thus each synthesized case record consists of NP+2 lines or unit records. The format is given in Table 10.

This tape was obtained by conversion of the binary output tape actually written by the program on the CDC 7600 computer at LBL. On the binary tape, each case appears as a single logical record, with no physical demarcation of individual lines. The first eight words are as listed in Table 10 for the formatted tape: all are type INTEGER except the sampling weights (words 5 and 6) which are type REAL . The census lines are coded as 10 words each, so that the case record consists of 10*NP + 18 words. A binary case record can be read by a FORTRAN program using, for example,

```
READ(1) ISEQ,IOZ,IDZ,IC,S1,S2,IP,NR,((IX(I,J),I=1,10),J=1,NR)
```

Then IX(1,1),...,IX(10,1) will contain the housing line and, if IP is not zero, IX(1,IP+1),...,IX(10,IP+1) will contain the person line of the selected worker.

TABLE 10. Format of Synthetic Sample Case Record

(formatted version of SYNSAM tape)

| Line | Variable | Format | Description |
|---|---|---|---|
| 1 | ISEQ | I5 | Case sequence number |
| | IOZ | I3 | Origin zone |
| | IDZ | I3 | Destination zone (set to zero if no workers in the household) |
| | IC | I4 | Socioeconomic cell number |
| | S1 | F8.6 | Work trip sampling weight |
| | S2 | F8.6 | Non-work trip sampling weight |
| | IP | I2 | Serial number of person record of the selected worker in this household (zero if no workers in the household) |
| | NR | I2 | Number of census lines to follow (NR = NP + 1) |
| 2 | | (*) | Household line |
| 3 | | (*) | Person line for person 1 |
| . | | . | |
| . | | . | |
| . | | . | |
| NR + 1 | | (*) | Person line for person (NR - 1) |

(*) See the Census Bureau documentation of the 1970 Public Use Sample (SMSA, 15%) for details of these formats. Each household and person line consists of 120 characters representing about 80 tabulations.

## REFERENCES

Bishop, T., S. Fienberg, and P. Holland (1975), Discrete Multivariate Analysis.

Darroch, J. (1962), "Interaction in Multi-factor Contingency Tables," J.R.S.S., Ser. B, Vol. 24, 251-263.

Deming, W. and F. Stephan (1940), "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known," Ann. Math. Stat., Vol. 11, 427-444.

Haberman, S. (1974), "Log-linear Fit for Contingency Tables," Appl. Stat., Vol. 21, 218-225.