2

Multidimensional Population Analysis with Incomplete Data

Frans Willekens

1. INTRODUCTION

In multidimensional demography, various *dimensions* of a demographic system are studied simultaneously. In multi-regional demography, the dimensions are age and region of residence. Demographic systems may also be defined along the dimensions age and marital status, age and labor force status, etc. For each dimension, several *states* or demographic categories may be distinguished, e.g., age groups of 5 years; marital states--never married, married, widowed, divorced. The states alive and death may also be considered to constitute a

**This paper is part of a research project toward the development of a methodology for inferring detailed migration patterns from incomplete data; the project is carried out with financial assistance of the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).*

dimension. In this paper, it is not done however, because deaths are calculated as residuals, i.e., as persons who do not move on to any of the states in the system.

The application of multidimensional demography puts a heavy burden on data availability. Data requirement is a major drawback of multidimensional analysis. The opportunities of advanced techniques of demographic inquiry can only be fully explored if data are abundantly available or if adequate procedures for estimating missing data can be used. Recently, statistical offices throughout the developed world have started collection and tabulation, on a regular basis, of necessary data for multidimensional population analysis or are advised to do so. It is particularly interesting to follow discussions in statistical offices on whether it is sufficient to collect information on sizes and compositions of major subpopulations, or whether it is necessary to quantify the flows among relevant categories of the subpopulations. In Section 2, we review some positive developments on the data collection scene. Although the steps taken toward the solution of the data problem are encouraging, one has still a long way to go. Even the data collected may not be adequate for multidimensional analysis because of deficiencies in sample frame or in sample sizes. The available data must therefore be supplemented frequently by estimates.

This chapter addresses the problem of estimating flow data that are required for multidimensional population analysis. The estimation of flows or transitions is a relatively new problem in demography. It has however been investigated for many years in regional science (migration), transportation science (traffic between nodes on a transportation network),

and economics (interindustry transactions in an input-output table). Statisticians have also devoted attention to this problem in information theory, in Bayesian statistics, and in contingency table analysis. We shall draw on this research to propose a strategy for estimating missing data in multidimensional population analysis. The main feature of this strategy is to derive the value of each missing element in the flow matrix from the structure of the whole data set. This means that individual elements are not considered on their own, but only in connection with the other elements, i.e., as components of a major structure. Accounting provides a framework for this structural representation.¹ The proposed strategy consists of five steps:

(a) Set up the accounting frame. Determine the dimensions (classifications) of the population system and the states (categories) in each dimension. The accounts studied in this paper are multidimensional cross classifications. As soon as the accounting frame is imposed, the location of the cells in the account tells something about the characteristics of the individuals falling into them. For instance, in a three-dimensional account individuals in a specific cell have a characteristic in common with individuals in all cells of the same row, another characteristic is common with all individuals in cells of the same column, and still another characteristic in common with all individuals in cells of the same layer.

¹A different, but not unrelated framework is provided by postulating an appropriate stochastic model. This approach is illustrated in Section 3.2 and is fully developed in de Jong (1981).

(b) Develop a model of the data set in the account. In order to be able to consider individual elements in connection with other elements, the structural relationships in the data are represented by a parametric model. Various models of data structures may be imagined, but we shall limit ourselves to a log-linear model. It can be shown that models developed in the disciplines mentioned above are particular formulations of the log-linear model, as conceived in research on patterns of association in contingency tables.

(c) Enter the available data into the account, i.e., fill in the account as far as possible. In general, we do not know individual cell counts but only marginal totals. If some of the cell counts are observed, enter them into the account. Other prior information is listed separately.

(d) Determine the parameter values of the parametric model on the basis of the different types of prior information, supplemented by hypotheses about certain structural relationships in the data to be estimated. A hypothesis, frequently used when only aggregated data are available, is the assumption of independence between some of the variables. Model fitting and testing for independence are equivalent in the modeling perspective adopted in this chapter. This equivalence allows a remarkably simple and transparent approach to estimating missing elements in the account.

(e) Apply the model to infer the values of the missing elements.

In implementing this five-step procedure, methodological and practical problems may arise. Some of the potential problems are dealt with in this chapter. Section 2 focuses on data

requirement and data availability. An illustrative overview is given of types of flow data that are available and of sources that may provide relevant data for multidimensional analysis. The overview is highly incomplete, but it demonstrates that it is worthwhile to search for additional and maybe uncommon data sources in order to fill the account as far as possible.

Section 3 discusses the log-linear model. Two model formulations are represented: the additive model is popular in interaction analyses; the multiplicative model may however be more appropriate in dealing with estimation problems. This section also reveals what information is sufficient to estimate the log-linear model parameters. These sufficiency conditions directly lead to the main theme of this paper, namely, the computation of expected cell counts from whatever data are available.

Different hypothetical situations of data availability are considered in Section 4:

- (i) marginal totals only;
- (ii) marginal totals, supplemented by preliminary estimates of the cell counts
- (iii) marginal totals and a few cell counts are given (i.e., exactly known).

These and related conditions of prior knowledge may be treated in a unified way, and a single multiproportional adjustment procedure for estimating the missing elements may be applied.

This general estimation technique is based on the log-linear analysis of the data sets and encompasses several approaches, found in the literature, to infer entries of an

n-dimensional table from available data. Gravity modeling, entropy maximization, information gain minimization, and other popular estimation methods are equivalent to multiproportional adjustment. This equivalence indicates some interesting directions for further research aiming at the improvement of estimation techniques.

The techniques presented in this chapter, as well as some equivalent procedures, are applied to real data in Section 5. Two fields of application are selected: the estimation of social mobility tables and of migration tables.

2. DATA REQUIREMENTS AND DATA AVAILABILITY

Multidimensional demographic analysis requires gross flow data. Conventional demographic analysis of populations subdivided into subpopulations focuses on the size of each subpopulation and on changes in size. This stock perspective makes use of prevalence rates, which indicate the extent to which a particular characteristic is prevalent in the population (e.g., labor force participation rate, proportion of married population). In multidimensional analysis, population size is of secondary importance. The emphasis is on flows, i.e., passages among the various states; the magnitude of each subpopulation is merely an outcome of an initial condition and a flow mechanism. The rationale for adopting the flow perspective is that the dynamics underlying population change can more readily be represented. Entrances to and exits from each subpopulation are considered explicitly. The adoption of the flow perspective requires the existence of flow data. These

data may be arranged in an accounting framework. The accounting framework is not only a convenient data-representation scheme, it is also a useful basis for the integration of data-estimation procedures. The accounting framework for multidimensional data representation and estimation is presented in Section 2.1. Section 2.2 gives an illustrative review of data sources for multidimensional population analysis and shows that the data problem is not all that bad. Several data sources with the right information have not yet been touched.

2.1. DATA REQUIREMENTS: AN ACCOUNTING FRAMEWORK

Multidimensional analysis requires age-specific flows among the various states considered in the analysis. The data may be arranged to constitute a multidimensional contingency table. This table represents the accounting framework in which the available information may be arranged and the estimation problem may be defined. In this chapter, we limit ourselves to a two-dimensional population system. One of the dimensions will be age. The findings may be generalized to any number of dimensions or, alternatively, a population system of a higher dimension can be reduced to two dimensions by increasing the number of states or categories in one or more dimensions (e.g., each of the four marital states may be sectioned by sex yielding eight states along the sex-marital status dimension).

For each age group, the number of passages from each state to any of the other states must be known. This information may be arranged in layers of two-way tables (Fig. 1). Let age (layer) be denoted by k , state of origin (row) by i , and state

Status of origin	Status of destination				
	1	2	3	Total	
	1	m_{11k}	m_{12k}	m_{13k}	$m_{1.k}$
	2	m_{21k}	m_{22k}	m_{23k}	$m_{2.k}$
	3	m_{31k}	m_{32k}	m_{33k}	$m_{3.k}$
Total	$m_{.1k}$	$m_{.2k}$	$m_{.3k}$	$m_{..k}$	

Fig. 1. Account of passages for age group k .

of destination (column) by j . There are L age groups (layers), R origins (rows), and C destinations (columns) ($R = C$). The total number of cells in the account is therefore $L \times R \times C$. Let K , I , and J represent the index sets of, respectively, k , i , and j :

$$K = \{1, 2, \dots, k, \dots, L\},$$

$$I = \{1, 2, \dots, i, \dots, R\},$$

$$J = \{1, 2, \dots, j, \dots, C\}.$$

Some passages or aggregates of passages may be fixed, because they are actually known or observed, while others must be estimated. Let S be the set of cells in a multiway array or account that must be estimated. Formally, $S = \{(i, j, k) \mid \text{passage from } i \text{ to } j \text{ by category } k \text{ is possible, and not fixed}\}$. If the (i, j, k) -cell lies in S , we write $(i, j, k) \in S$. Reference will be made to four types of arrays: observed array $\{x_{ijk}\}$ containing the observed counts of the number of persons in category k passing from i to j ; array of expected counts

$\{m_{ijk}\}$; prior array or array of preliminary estimates $\{m_{ijk}^0\}$ and array of maximum likelihood estimates (MLEs) of the expected counts, $\{\hat{m}_{ijk}\}$.

In practical applications, the array $\{x_{ijk}\}$ is not known, except for elements $(i,j,k) \in S$, if any. The main use of an observed array lies in validity analyses of estimation methods.

Aggregations over the index sets I , J , or K are denoted by dots. For instance, the bivariate marginal $x_{.jk}$ is the marginal total of x_{ijk} over all $i \in I$. The univariate marginal $x_{..k}$ is the sum of x_{ijk} over all $i \in I$ and $j \in J$. The overall total is $x_{...} = N$ and is equal to the total number of passages in the system. The marginal totals in the account are of particular relevance in multidimensional analysis with incomplete data since the information available of the passages generally is limited to aggregated flow values.

In this section, the term "passage" has not been defined. The definition very much depends on the way passages are measured and hence on the data collection system. In censuses and retrospective surveys, a passage is measured by comparing the status at the time of enumeration with the status at a prior date. In registration systems, however, each change in status is accounted for. Whatever passage definition is retained, the accounting framework is a useful strategy to integrate data availability and data estimation. Accounts also provide a logical connection between data and demographic models. Rees and Wilson (1977) and Rees (1980) elaborate on the advantages of accounting for demographic model building and propose some rules to designing accounts. The reader, interested in the problem of setting up an account, is

referred to the literature, in particular to Rees (1980); in this chapter, we assume the accounting framework to be given and focus on the modeling of the data in the account and on the estimation of missing elements.

2.2. ILLUSTRATIVE REVIEW OF SOURCES OF DATA FOR MULTIDIMENSIONAL POPULATION ANALYSIS

In this section, we list some data sources that provide highly relevant data for multidimensional analysis. This review is not exhaustive, its only purpose is to illustrate the kind of data that are available and the types of estimation problems associated with it. Data availability is a fuzzy concept. Data may have been collected but not tabulated, or they may have been tabulated but not published. In the latter case, the data are generally available on microfiches; in the first case, they may be tabulated on request. This review considers data for multiregional analysis, labor force analysis, and marital status analysis. Some new areas of application of multidimensional demography are also discussed.

A. Multiregional analysis. Censuses and registration systems are the main sources of migration data. In several European countries, any change of residence must be registered. Central statistical offices collect the information from the local administration to prepare migration statistics. The "movement card" contains information on origin and destination, as well as on certain demographic characteristics such as age. The census is the main source of migration data in most countries. Relevant data may be derived from answers to such questions as age, place of enumeration, place of birth, place

of previous residence, place of residence at a fixed prior date (1 or 5 yr ago), and duration of residence. An advantage of censuses is that they provide detailed information on migrant characteristics. A disadvantage, however, is that the information provided may not be up-to-date, because of the interval at which censuses are taken. Therefore, census data may be supplemented with migration information obtained through other sources such as household surveys and labor force surveys, which are also held at regular intervals. This poses the problem, however, of combining separate data sources; a problem that will be dealt with in this chapter.

B. Labor force analysis. Multidimensional tables of working life have been constructed for Denmark (Hoem and Fong, 1976; Willekens, 1980b) and for the United States (Schoen and Woodrow, 1980; Smith, 1980). Hoem and Fong use flow data generated by a special labor force survey in 1973-1974. The Schoen-Woodrow and Smith studies rely on Current Population Survey (CPS) data, provided by the Bureau of Census. Schoen and Woodrow give a detailed description of the data. Since January 1973, the CPS contains a retrospective question on the labor force status exactly one year ago. However, this question was only asked to persons aged 16 and over who are employed at the time of enumeration (about 60% of the total number of respondents). As a consequence, only part of the account can be filled with observed data. In the future, a more complete set of labor force flow data may become available in the U.S. The National Commission on Employment and Unemployment Statistics recently recommended that the CPS resume publication of gross flow data on occupational mobility and prepare

monthly gross flow data time series tapes for public use. It also recommended that the next census questionnaire include a question on occupation, industry, and place of residence one year ago (Stein, 1980).

A potentially very useful data source for multidimensional labor force analysis is the biannual Labor Force Survey held since 1973 in each member country of the European Community. This survey distinguishes eight labor force statuses (of which one is employed) and contains for each status a retrospective question on the labor force status one year ago (since 1977, also the status two years ago is being asked). The survey, which is set up on a comparable basis in each country, provides rich demographic and socioeconomic information on the respondents. Care must however be given to extract only the information that is supported by the sample size. The Dutch National Program for Demographic Research initiated a project aiming at the development of multidimensional tables of working life for the Netherlands on the basis of the transition data provided by the Labor Force Survey. In a later stage, it will be investigated whether these data may also be used to project the labor force.

C. Nuptiality analysis. Flow data on marital formation and dissolution are more readily available than data for other applications of multidimensional analysis. The registration of changes in marital status is common in most countries. The Belgian and Dutch data situations are discussed in Willekens *et al.* (1979) and Koesoebjono (1981), respectively. Schoen and Nelson (1974) provide some insight into the kind of data on marital change available in the U.S.

D. Education analysis and other possible applications of multidimensional demography. Multiregional analysis, labor force, and nuptiality analysis, are only a few of the possibly many situations in which multidimensional demographic techniques may fruitfully be applied. In fact, any demographic investigation of renewable events may benefit from application of the increment-decrement life table model and hence from multidimensional demography. The growing number of retrospective surveys may furnish the required input data. Education analysis is one new possible field of application of multidimensional demography. In The Netherlands, data on flows to, from, and within the educational system were collected by the Central Bureau of Statistics in a retrospective survey of pupils in 1978. The survey inquired about type and level of education one year ago. The published flow matrix covers a large set of educational categories; the age dimension has however been neglected, but can easily be recovered. Other topics of analysis in which the application of multidimensional demographic techniques may lead to greater insight include the study of social mobility, family planning, and social security participation. For several of these studies, flow data may be available from retrospective surveys. If data availability is inadequate, however, multidimensional demography may still be considered. Appropriate estimation methods may infer the required data from the available knowledge. Such techniques are the subject of this chapter.

3. MODELING THE DATA IN THE ACCOUNT

The investigation of large data sets becomes relatively simple by fitting models to the data. During the past decade, the structural analysis of contingency tables attracted considerable interest, and the results of this research are well documented (Bishop *et al.*, 1975; Goodman, 1978; Gokhale and Kullback, 1978; Haberman, 1979). Analytical techniques, originally developed to identify patterns of associations among several categorical variables, may fruitfully be applied for estimation purposes.² The log-linear model is one of them. It is part of a class of generalized linear models, describing cell counts in terms of marginal totals and of interactions between cross-classified variables.

In Section 3.1, the log-linear model is presented; it is shown that it is an efficient tool for structural analysis of categorical data. Section 3.2 gives a simple rule for determining the necessary information to estimate the model parameters.

3.1. THE LOG-LINEAR MODEL

The log-linear model is not unknown in demography. Recently, an increasing number of authors is adopting this

²In mathematical statistics, a distinction is made between "estimation" and "prediction." Estimation refers to model parameters, while prediction relates to the outcomes (of random variables) obtained by applying or postulating a particular model. In this chapter, no distinction is made because in the demographic literature "migration prediction" has a well-defined but different meaning. For the statistical theory of prediction as applied to migration, the reader is referred to de Jong (1981).

modeling perspective to study dependencies among cross-tabulated demographic variables (Little, 1978, 1980; Little and Pullum, 1979; Hobcraft, 1978; Clogg, 1978, 1979; Fienberg and Mason, 1978).

The current formulation of the log-linear model is due to Birch (1963) and resembles the analysis-of-variance model. A neat review is by Payne (1977). Table 1 shows the model for the expected cell counts in two equivalent forms. The additive formulation is most popular since it closely resembles the analysis-of-variance paradigm. The multiplicative formulation is however convenient to solve estimation problems since it directly relates to conventional estimation models for flow data. This relationship to conventional models simplifies the interpretation of the model parameters in terms of available data. In addition, algorithms developed in regional science and transportation science may fruitfully be applied to solve estimation problems in multidimensional population analysis.

In both the additive and the multiplicative formulations, there are eight terms in the model. The number of terms depends on the dimension of the account and is unrelated to the number of cells or, equivalently, to the number of states or categories along the dimensions. The number of parameter-values, however, does depend on the number of cells in the account. In models (1) and (11), there are as many independent parameter-values as there are cells in the account. The model is therefore referred to as a *saturated log-linear model*. Each parameter in the model represents a particular structural effect on m_{ijk} . According to the log-linear model, the expected cell count is the sum of various effects. The overall

Figure 1. THE LOG-LINEAR MODEL

Multiplicative formulation

$$\text{Model } m_{ijk} = w \overset{A}{w_i} \overset{B}{w_j} \overset{C}{w_k} \overset{AB}{w_{ij}} \overset{AC}{w_{ik}} \overset{BC}{w_{jk}} \overset{ABC}{w_{ijk}} \quad (1)$$

$$\text{Overall mean effect } w = \left[\prod_{i,j,k} m_{ijk} \right]^{\frac{1}{RCL}} \quad (2)$$

$$\text{Main effects } \overset{A}{w_i} = \frac{1}{w} \left[\prod_{j,k} m_{ijk} \right]^{\frac{1}{CL}} \quad (3)$$

$\overset{B}{w_j}$ and $\overset{C}{w_k}$: analogous

First-order interaction effects (two-way or pairwise interaction)

$$\overset{AB}{w_{ij}} = \frac{1}{z} \left[\prod_k m_{ijk} \right]^{\frac{1}{L}} \quad (4)$$

$$\text{with } z = w \overset{A}{w_i} \overset{B}{w_j} \overset{C}{w_k}$$

$$\overset{AC}{w_{ik}} \text{ and } \overset{BC}{w_{jk}}: \text{ analogous}$$

Second-order interaction effects (three-way interaction)

$$\overset{ABC}{w_{ijk}} = \frac{1}{z'} m_{ijk}, \quad (5)$$

$$\text{with } z' = w \overset{A}{w_i} \overset{B}{w_j} \overset{C}{w_k} \overset{AB}{w_{ij}} \overset{AC}{w_{ik}} \overset{BC}{w_{jk}}$$

$$\text{Constraints } \prod_i \overset{A}{w_i} = \prod_j \overset{B}{w_j} = \prod_k \overset{C}{w_k} = 1 \quad (6)$$

$$\prod_i \overset{AB}{w_{ij}} = \prod_j \overset{AB}{w_{ij}} = \prod_i \overset{AC}{w_{ik}} = \prod_k \overset{AC}{w_{ik}} = \prod_j \overset{BC}{w_{jk}} = \prod_k \overset{BC}{w_{jk}} = 1 \quad (7)$$

$$\prod_i \overset{ABC}{w_{ijk}} = \prod_j \overset{ABC}{w_{ijk}} = \prod_k \overset{ABC}{w_{ijk}} = 1 \quad (8)$$

$$w = \exp u \quad (9)$$

$$\begin{aligned} \overset{A}{w_i} &= \exp u_i^A \\ \overset{AB}{w_{ij}} &= \exp u_{ij}^{AB} \\ \overset{ABC}{w_{ijk}} &= \exp u_{ijk}^{ABC} \end{aligned} \quad (10)$$

Additive formulation

$$\ln m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC} \quad (11)$$

$$u = \frac{1}{RCL} \sum_{i,j,k} \ln m_{ijk} \quad (12)$$

$$u_i^A = \frac{1}{CL} \sum_{j,k} \ln m_{ijk} - u \quad (13)$$

u_j^B and u_k^C : analogous

$$u_{ij}^{AB} = \frac{1}{L} \sum_k \ln m_{ijk} - c \quad (14)$$

with $c = u + u_i^A + u_j^B + u_k^C$

u_{ik}^{AC} and u_{jk}^{BC} : analogous

$$u_{ijk}^{ABC} = \ln m_{ijk} - c', \quad (15)$$

with $c' = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC}$

$$\sum_i u_i^A = \sum_j u_j^B = \sum_k u_k^C = 0 \quad (16)$$

$$\sum_i u_{ij}^{AB} = \sum_j u_{ij}^{AB} = \sum_i u_{ik}^{AC} = \sum_k u_{ik}^{AC} = \sum_j u_{jk}^{BC} = \sum_k u_{jk}^{BC} = 0 \quad (17)$$

$$\sum_i u_{ijk}^{ABC} = \sum_j u_{ijk}^{ABC} = \sum_k u_{ijk}^{ABC} = 0 \quad (18)$$

$$u = \ln w \quad (19)$$

$$u_i^A = \ln w_i^A$$

$$u_{ij}^{AB} = \ln w_{ij}^{AB}$$

$$u_{ijk}^{ABC} = \ln w_{ijk}^{ABC} \quad (20)$$

effect is a size-effect; it is the geometric mean of all cell counts. The main effects denote the effects on m_{ijk} of relative size differences between the various univariate marginals. For instance, w_k^C and u_k^C are the effects of the age classification on the number of passages m_{ijk} . Everything else equal, large age groups result in large values of m_{ijk} . The age effect is the ratio between the geometric mean of the k -th layer and the overall geometric mean. By comparing various geometric means, we can determine the range of effects exerted upon m_{ijk} . For example, to determine whether the age effect differs with the state of origin i , it suffices to calculate w_{ik}^{AC} or u_{ik}^{AC} . A nonzero value of u_{ik}^{AC} means the existence of an interaction between age and origin. Note that the pattern of interaction determined in this way represents the average interaction over all AC-tables (for all possible destination regions, i.e., j values of variable B). The pattern may differ for each level of B, resulting in a nonzero u_{ijk}^{ABC} . If $u_{ijk}^{ABC} \neq 0$, then the interaction between pairs of A, B, and C must also be nonzero. The principle that for any nonzero u-term its lower-order relatives must also be nonzero is referred to as the hierarchy principle (Bishop *et al.*, 1975, p. 34). Conversely, if any u-term is set equal to zero, its higher-order relatives must also be zero. In this chapter, only *hierarchical log-linear models* are considered.

By introducing the log-linear model, we have transformed the problem of estimating the cell counts into a problem of parameter-estimation, i.e., of quantifying the various effects. The effects are determined on the basis of available data augmented by assumption-making.

The following conclusion is the basis for the derivation of the estimation procedure. *To estimate the cell values in a multidimensional account, we must quantify the interaction effects. Hence data estimation is closely related to hypothesis testing.* If all data are available, i.e., if the array $\{x_{ijk}\}$ is known, then all parameter-values may be derived from the data. In Table 4 of Section 5, the parameters are shown of the log-linear model of a set of social mobility data. It can easily be seen that the saturated log-linear model replicates the observed data exactly, i.e.,

$$m_{ijk} = x_{ijk}.$$

In practical applications, $\{x_{ijk}\}$ is not known and the parameters must be calculated from whatever prior information exists. If certain parameters can not be estimated from available data, their value is set equal to zero, and the associated interaction effect is assumed to be absent. A log-linear model with certain terms absent is an *unsaturated model*. Section 3.2 describes an integrated approach to parameter estimation. It should be kept in mind that by estimating a log-linear model parameter, we are imposing a particular interaction pattern on the $\{m_{ijk}\}$ -array. Conversely, if we want the estimates to express that certain variables depend on one another in a particular way, we must introduce this through appropriate values of the relevant model parameters.

3.2. SUFFICIENT STATISTIC FOR OBTAINING CELL ESTIMATES

First we determine what information is required for estimating the log-linear model parameters. To do this, we consider the account to be the outcome of a simple multinomial sampling scheme (sample in which the total sample size is fixed, say N , and each cell has an independent Poisson distribution). The account denotes then a multinomial distribution $M(N; m_{ijk}/m_{...})$ with probability density function (Fisher, 1922; Bishop et al., 1975, p. 63):

$$P(X_{ijk} = x_{ijk} \text{ for all } i, j, k | X_{...} = N) \\ = \frac{N!}{\prod_{ijk} x_{ijk}!} \prod_{ijk} \left(\frac{m_{ijk}}{m_{...}} \right)^{x_{ijk}}. \quad (21)$$

The log-likelihood of the multinomial is

$$\ln \left\{ \frac{N!}{\prod_{i,j,k} x_{ijk}!} \right\} + \sum_{i,j,k} x_{ijk} \ln m_{ijk} - N \ln m_{...}. \quad (22)$$

In maximizing the log-likelihood, under the assumption of $\sum_{ijk} (m_{ijk}/m_{...}) = 1$, we may disregard the first and third term and only consider the remaining term, the kernel of the log-likelihood function. The product-form of this kernel

$$W = \prod_{ijk} (m_{ijk})^{x_{ijk}} \quad (23)$$

represents the expected number of individuals falling into cell (i, j, k) if x_{ijk} members of the population of size N are randomly selected. The quantity W/N is, therefore, the probability that a randomly selected individual falls into the (i, j, k) -cell.

Substituting $\ln m_{ijk}$ in the kernel for the log-linear model (11) gives:

$$\begin{aligned}
 \sum_{i,j,k} x_{ijk} \ln m_{ijk} = & Nu + \sum_i x_{i..} u_i^A + \sum_j x_{.j.} u_j^B \\
 & + \sum_k x_{..k} u_k^C + \sum_{i,j} x_{ij.} u_{ij}^{AB} \\
 & + \sum_{i,k} x_{i.k} u_{ik}^{AB} + \sum_{j,k} x_{.jk} u_{jk}^{BC} \\
 & + \sum_{i,j,k} x_{ijk} u_{ijk}^{ABC} .
 \end{aligned} \tag{24}$$

The sufficient statistic for estimating the parameters of the log-linear model consists of the x-terms adjacent to the unknown parameters. For instance, to determine the column effect u_j^B , we need to know the total $x_{.j.}$; and to determine the interaction between A and B, the marginal total $x_{ij.}$ must be known. Note that the interaction derived from $x_{ij.}$ is an average of the interaction patterns at various k-levels of the third variable C.

Knowledge of the bivariate marginal $x_{ij.}$ implies knowledge of the univariate marginals $x_{i..}$ and $x_{.j.}$ necessary to determine u_i^A and u_j^B , respectively. Hence we can define a minimal sufficient statistic (in this case $x_{ij.}$). In practice, the minimal sufficient statistic can be obtained by inspection of the log-linear model.

Expression (24) relates the log-linear model to the likelihood function. Birch (1963) has shown that there is a unique set of elementary cell-estimates that

- (i) satisfies the conditions imposed by the form of the parametric model (log-linear);
- (ii) satisfies the constraints that the marginal totals of the estimates m_{ijk} correspond to the given marginal totals ($x_{ij.}$, $x_{.j.}$, etc.); and
- (iii) maximizes the likelihood function.

Hence, by applying the log-linear model, maximum likelihood estimates $\hat{m}_{ijk}$ of the expected counts m_{ijk} may be derived from the sufficient statistic alone. Formulated differently and more in line with the main theme of this chapter, application of the log-linear model to available data yields estimates that not only are consistent with what is known, but that also have a maximum likelihood of occurrence. The informed reader will notice the equivalence between these features and the characteristics of the estimates obtained by applying entropy maximization techniques, which have been developed by Wilson (1970) in the field of regional science. The equivalence between entropy maximization and maximum likelihood estimation is formally demonstrated by Batty and Mackie (1972) and Willekens (1980a).

4. PARAMETER ESTIMATION FROM AVAILABLE DATA

To estimate missing elements in the multidimensional account, a modeling approach was suggested and the log-linear model was selected as an adequate representation of the data. The functional form of the parametric model, underlying the estimation procedure, is therefore fixed. The sufficient sta-

tistic for estimating the model parameters consists of the x -terms adjacent to the unknown parameters. In this section, we review a procedure to derive parameter-values for this model from incomplete data. First, we assume that the prior information is limited to marginal totals only. Next, it will be shown how other types of prior information may be introduced. Whatever prior data are used, the estimation strategy is the same: prior knowledge yields estimates of log-linear model parameters and hence imposes a structure on the expected counts $\{m_{ijk}\}$. The better we can describe the structural relationship between the cross-classified variables, the better the estimates will be. If we are unable to obtain some parameter-values because of lack of information, observed or approximated, on the adjacent x -terms, the values are set equal to zero, implying the absence of the interaction pattern they represent.

4.1. METHODS OF OBTAINING CELL ESTIMATES FROM MARGINAL TOTALS ONLY

The MLEs we are looking for satisfy the log-linear model and the marginal constraints. The model and the constraints form a system of equations, the maximum likelihood equations, the solution of which yields the MLEs. In some cases, the MLE-equations have a closed-form solution.

A. Closed-Form Expressions

The estimates may be expressed in closed form if the known marginal totals satisfy particular conditions. For example, if the available information is limited to $x_{i..}$, $x_{.j.}$, and $x_{..k}$, then the MLEs $\hat{m}_{ijk}$ are the solution of the following system of equations:

Model equations:

$$m_{ijk} = w w_i^A w_j^B w_k^C$$

or

$$\ln m_{ijk} = u + u_i^A + u_j^B + u_k^C \quad (25)$$

with the w - and the u -parameters satisfying (6) and (16), respectively.

Minimal sufficient statistic equations:

$$\sum_{j,k} \hat{m}_{ijk} = x_{i..} \quad , \quad (26)$$

$$\sum_{i,k} \hat{m}_{ijk} = x_{.j.} \quad , \quad (27)$$

$$\sum_{i,j} \hat{m}_{ijk} = x_{..k} \quad , \quad (28)$$

$$\sum_{i,j,k} \hat{m}_{ijk} = x_{...} = N \quad . \quad (29)$$

The cell estimates for model (25) and the sufficient statistic consisting of (26) to (29) are

$$\hat{m}_{ijk} = \frac{x_{i..} x_{.j.} x_{..k}}{N^2} \quad .$$

This formula provides a closed-form expression of MLEs in terms of sufficient statistics. It is the simplest illustration of the multidimensional estimation problem. Willekens *et al.* (1979) label this case the 3E (edge)-problem since the known information may be arranged on the edges of a box, the contents of which must be estimated. Several other closed-form estimates in three dimensions exist. They may be derived

by solving the appropriate set of maximum likelihood equations. Table 2 summarizes the results. The available data are represented as a set of univariate and/or bivariate marginals. In a multiregional population system, for instance, a case could arise in which the migration pattern is known for the total population and only a single age structure of migrants is given. This prior information could be arranged on one face and one edge of a box and hence the label IFE (face-edge)-problem may be used. The MLE procedure reduces to applying the age composition $\frac{1}{N} x_{..k}$ to each of the elements of the migration matrix $x_{ij.}$. The application of a single age composition implies that in the m_{ijk} array, age is independent of region of origin and of region of destination. Although an unrealistic assumption, it is enforced on the model because of limited prior information. Since we do not have sufficient statistics to estimate all the parameters of the log-linear model, we postulate that some parameters are zero, with the consequence that the interaction effects they represent are absent. In the next section we shall see that even in the absence of "hard" data, we can impose some structure on the $\{m_{ijk}\}$ array by using "softer" information such as data collected at some prior date, expert opinion, relative measures obtained in sample surveys, etc.

The existence of closed-form expressions for the MLEs is very convenient in multidimensional analysis with incomplete data. Bishop *et al.* (1975, pp. 76-82) provide a few rules for detecting the existence of direct estimates. The interesting feature of these rules is that they apply to accounts of any dimension. The main idea is to delete the redundant configurations of available data or overlapping subconfigurations. If

TABLE 2. MLEs in Three-Dimensional Accounts

Case	Available data	Log-linear model	Closed-form cell estimates	Interpretation of log-linear
3E	{A, B, C}	$m_{ijk} = w \frac{A}{i} \frac{B}{j} \frac{C}{k} w_{ijk}$	$\hat{m}_{ijk} = \frac{1}{N^2} x_{i..} x_{.j.} x_{..k}$	Mutual independence (variables A, B, and C are independent)
IFE	{AB, C}	$m_{ijk} = w \frac{A}{i} \frac{B}{j} \frac{C}{k} \frac{AB}{ij} w_{ijk}$	$\hat{m}_{ijk} = \frac{1}{N} x_{ij.} x_{..k}$	Multiple independence (joint variable AB is independent of C)
2F	{AB, BC}	$m_{ijk} = w \frac{A}{i} \frac{B}{j} \frac{C}{k} \frac{AB}{ij} \frac{BC}{jk} w_{ijk}$	$\hat{m}_{ijk} = x_{ij.} x_{.jk} / x_{.j.}$	Conditional independence (A independent of C, given B)
3F	{AB, AC, BC}	$m_{ijk} = w \frac{A}{i} \frac{B}{j} \frac{C}{k} \frac{AB}{ij} \frac{BC}{jk} \frac{AC}{ik} w_{ijk}$	No closed-form solution	Pairwise association (each two-way interaction is independent of level of third variable)
CI	{ABC}	$m_{ijk} = w \frac{A}{i} \frac{B}{j} \frac{C}{k} \frac{AB}{ij} \frac{BC}{jk} \frac{AC}{ik} \frac{ABC}{ijk} w_{ijk}$	$\hat{m}_{ijk} = x_{ijk}$	Three-way interaction (association between every pair of variables varies with level of third variable)

no more than two configurations remain, closed-form estimates exist. In general (for any dimension), it can be shown that this statement implies that at least one two-factor effect must be absent for direct estimates to exist. The general form of the direct estimates is predictable: the numerator has entries from each sufficient configuration; the denominator has entries from redundant configurations caused by overlapping; terms in powers of N ensure the right order of magnitude. The overlapping is illustrated in the 2F-case: the subscript j appears in both the AB and the BC configurations.

B. Iterative Fitting by Multiproportional Adjustment

To derive the MLEs, the 3F-problem requires a minimal sufficient statistic, consisting of three bivariate marginals. No closed-form solution exists, and iterative fitting of the sufficient configurations or prior data is the only way out. To start the procedure, we may choose any set of preliminary estimates that do not exhibit a three-factor effect ($u_{ijk}^{ABC} = 0$). For instance, the uniform distribution satisfies this condition, hence a convenient starting value is $m_{ijk}^{(0)} = 1$ for all i , j , and k .

The iterative algorithm goes as follows:

Step 0: $s = 0$

Step 1: Proportional adjustment along the C dimension

$$\hat{m}_{ijk}^{(3s+1)} = \hat{m}_{ijk}^{(3s)} \frac{x_{ij.}}{\sum_k \hat{m}_{ijk}^{(3s)}} .$$

Step 2: Proportional adjustment along the A dimension

$$\hat{m}_{ijk}^{(3s+2)} = \hat{m}_{ijk}^{(3s+1)} \frac{x_{.jk}}{\sum_i \hat{m}_{ijk}^{(3s+1)}} .$$

Step 3: Proportional adjustment along the B dimension

$$\hat{m}_{ijk}^{(3s+3)} = \hat{m}_{ijk}^{(3s+2)} \frac{x_{i.k}}{\sum_j \hat{m}_{ijk}^{(3s+2)}} .$$

If the stopping criterion

$$\left| \frac{\hat{m}_{ijk}^{(3s+3)}}{\hat{m}_{ijk}^{(3s+2)}} - 1 \right| \leq \epsilon$$

for every i , j , and k is satisfied, then stop the iteration; otherwise $s = s + 1$ and go to step 1.

The algorithm is a special variant of a more general algorithm discussed in Section 4.2. This method of successive proportional adjustment is known under various names. It was originally developed by Bartlett in 1935, who called it the "no second-order interaction model." In contingency table analysis, it became known as the iterative proportional fitting (IPF) method (see, e.g., Bishop *et al.*, 1975, pp. 83-97). In this chapter, the algorithm is referred to as multiproportional adjustment (MAD) algorithm, for reasons that will become clear later. Applying the Rockafellar decomposition principle, Willekens *et al.* (1979) show that this algorithm may be derived from a maximization of the entropy function

$$\sum_{i,j,k} m_{ijk} \ln m_{ijk} , \tag{30}$$

subject to the bivariate constraints $\{AB, AC, BC\}$ and that the method converges to a unique solution (a number of other proofs of convergence of IPF exist; see references in Bishop *et al.*, 1975, p. 85). They demonstrate that the IPF algorithm is equivalent to the direct primal algorithm of the nonlinear mathematical programming problem. In addition, the authors derive an algorithm, based on the dual formulation of the programming problem. The advantage of this dual formulation is that it relates more directly to the parameters of the log-linear model. The dual will be discussed in the next section.

To show that MAD does not introduce a third-order effect, we may rewrite the estimates as a product of functions in two variables only:

$$\hat{m}_{ijk} = f_1(i,j) \cdot f_2(j,k) \cdot f_3(i,k), \quad (31)$$

with

$$f_1(i,j) = \prod_s \frac{\hat{m}_{ij.}}{\sum_k \hat{m}_{ijk}^{(3s)}},$$

$$f_2(j,k) = \prod_s \frac{\hat{m}_{.jk}}{\sum_i \hat{m}_{ijk}^{(3s+1)}},$$

$$f_3(i,k) = \prod_s \frac{\hat{m}_{i.k}}{\sum_j \hat{m}_{ijk}^{(3s+2)}},$$

where s denotes the iteration. Equivalent algorithms to calculate the bivariate functions of (31) are presented by Chilton and Poet (1973) and Caussinus and Thelot (1976).

4.2. METHODS OF OBTAINING CELL ESTIMATES FROM MARGINAL TOTALS SUPPLEMENTED BY OTHER PRIOR INFORMATION

The x-terms adjacent to the unknown parameters are sufficient for estimating the parameters of the log-linear model [see expression (24)]. In the previous section we assumed the parameters to be zero and the interaction pattern they represent to be absent, if the required x-term was not available. In particular, we were unable to assign any value to the u^{ABC} -term since the individual cell counts x_{ijk} were unknown. If some x-terms are not available, the parameter estimates may be derived from other sources of prior information. A combination of different data sources may permit the derivation of adequate estimates for log-linear model parameters and for cell counts. We shall call *main data source* to set of known x-terms; *auxiliary data sources* is (are) the prior information from which the model parameters are derived that cannot directly be obtained from the x-terms. Prior information may come in various ways. For instance, detailed mobility data may be lacking for one country, but may exist for another country. If the two countries are similar, then log-linear model parameters obtained for one country may be applied to derive mobility estimates for the other country. Analogously, similar data sources of different periods may be combined. Censuses may have the required detailed information for multidimensional population analysis, but the statistics may be outdated. A combination of census information and more recent data may yield an adequate data base for multidimensional analysis. Sometimes, expert opinion, intuition, and common sense may be introduced to increase the quality of the data base. We may know in advance

that certain transitions are impossible or must take on certain values. For instance, reentrance into the state of never-married is impossible. In some cases, certain transitions are disregarded: in multiregional analysis, intraregional migrations are generally left out of consideration because they do not affect population redistribution. Fixed cell values arise when certain x_{ijk} -elements are observed. This is, for instance, illustrated by the data base used by Schoen and Woodrow (1980) to construct tables of working life. Labor force transitions were observed for a subgroup of the population only, namely, for those persons employed at the time of enumeration. The transitions made by persons in a different labor force category, had to be estimated.

The aim of this section is to show how an appropriate auxiliary data source may be generated if it does not exist, and how it may be combined with the main data source to yield accurate estimates for log-linear model parameters and cell counts. The main idea is that the parameters, which cannot be estimated from the main data source, are derived ("borrowed") from the auxiliary source.

4.2.1. *Combining Separate Data Sources*

The auxiliary data source is denoted by $\{x_{ijk}^0\}$ and it gives rise to the array or account of preliminary estimates $\{m_{ijk}^0\}$. The account is of the same dimension and magnitude as the $\{x_{ijk}\}$ -account. If all x_{ijk}^0 are observed, then $m_{ijk}^0 = x_{ijk}^0$. In this chapter, we assume that all x_{ijk}^0 are observed. Both the main and the auxiliary data sets may be modeled by the log-linear expressions:

$$\ln m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC}, \quad (32)$$

$$\ln m_{ijk}^0 = {}^0u + {}^0u_i^A + {}^0u_j^B + {}^0u_k^C + {}^0u_{ij}^{AB} + {}^0u_{ik}^{AC} + {}^0u_{jk}^{BC} + {}^0u_{ijk}^{ABC}, \quad (33)$$

and

$$\ln \frac{m_{ijk}}{m_{ijk}^0} = r_u + r_{u_i^A} + r_{u_j^B} + r_{u_k^C} + r_{u_{ij}^{AB}} + r_{u_{ik}^{AC}} + r_{u_{jk}^{BC}} + r_{u_{ijk}^{ABC}}, \quad (34)$$

with

$$u = u - {}^0u,$$

$$r_{u_i^A} = u_i^A - {}^0u_i^A, \quad r_{u_j^B} = u_j^B - {}^0u_j^B, \quad r_{u_k^C} = u_k^C - {}^0u_k^C,$$

$$r_{u_{ij}^{AB}} = u_{ij}^{AB} - {}^0u_{ij}^{AB}, \quad r_{u_{ik}^{AC}} = u_{ik}^{AC} - {}^0u_{ik}^{AC}, \quad r_{u_{jk}^{BC}} = u_{jk}^{BC} - {}^0u_{jk}^{BC},$$

$$r_{u_{ijk}^{ABC}} = u_{ijk}^{ABC} - {}^0u_{ijk}^{ABC}.$$

The multiplicative formulation of the log-linear model (34) is

$$m_{ijk} = m_{ijk}^0 r_w r_{w_i^A} r_{w_j^B} r_{w_k^C} r_{w_{ij}^{AB}} r_{w_{ik}^{AC}} r_{w_{jk}^{BC}} r_{w_{ijk}^{ABC}}, \quad (35)$$

where the superscript r denotes the ratio of the w -term in the log-linear model of the final estimates to the 0w -term in the log-linear model of the preliminary estimates. Assuming that the auxiliary account is completely observed ($m_{ijk}^0 = x_{ijk}^0$), the associated log-linear model is saturated and all the 0u -parameters, including the second-order interaction term, may be calculated from the preliminary estimates. Values of the u -parameters, on the other hand, are determined as follows:

- (i) for available x-terms, the associated u-parameters are calculated from the main data source;
- (ii) the u-parameters that cannot be determined from the main data source are set equal to the equivalent 0u -terms.

According to this procedure, interaction patterns between variables in the $\{m_{ijk}\}$ -account are derived from the main data source as far as possible; interaction patterns, which cannot be determined this way, are assumed to be equal to the interaction patterns observed in the auxiliary account. We may now answer an important question: How does the auxiliary data source contribute to the quality of the m_{ijk} -estimates? The auxiliary data set introduces interaction patterns into the final estimates that cannot be derived from the main data source. As a corollary, we may conclude that through a selection of an appropriate $\{m_{ijk}^0\}$ -array, combined with the right set of x-terms, we can impose any desired higher-order interaction patterns onto the $\{m_{ijk}\}$ -account. This approach may be of great use in multidimensional population analysis with incomplete data. The estimates obtained are MLEs or are approximations to MLEs depending on the data available (Haberman, 1979, pp. 519-540).

The $\{m_{ijk}\}$ -account that satisfies the conditions spelled out above, may be obtained without calculation of the log-linear model parameters as an intermediate step. Proportional adjustments of the cells of the $\{m_{ijk}^0\}$ -account to fit a set of predefined margins (x-terms) yield the appropriate estimates. If three bivariate marginal totals ($x_{ij.}$, $x_{i.k}$, $x_{.jk}$) are known, the algorithm closely resembles the Bartlett procedure, mentioned in the previous section and denoted as multipropor-

tional adjustment. Instead of starting with the uniform distribution, yielding a "no second-order interaction" model, the $\{m_{ijk}^0\}$ -array is used:

Step 0: $m_{ijk}^{(0)} = m_{ijk}^0$ for all $i, j, k \in S$

The algorithm converges to MLEs (Haberman, 1979, p. 540). In this 3F-case, the main effects and first-order interaction effects are derived from the given marginals; only the second-order interaction effects are "borrowed" from the $\{m_{ijk}^0\}$ -array, i.e., $u_{ijk}^{ABC} = {}^0u_{ijk}^{ABC}$ and, hence, $r_{u_{ijk}^{ABC}} = 0$. The log-linear model of this 3F-case is easily derived from (34) by deleting the second-order interaction term. The log-linear model is equivalent to the expression obtained by Willekens *et al.* (1979, p. 23) through maximizing the dual of the entropy function

$$\sum_{i,j,k} m_{ijk} \ln(m_{ijk}/m_{ijk}^0)$$

subject to the three sets of bivariate constraints that constitute the minimal sufficient statistic.

The solution of the dual entropy problem is

$$m_{ijk} = m_{ijk}^0 \exp[-(1 + \lambda_{ij} + v_{ik} + \xi_{jk})] , \quad (36)$$

with λ_{ij} , v_{ik} , and ξ_{jk} the dual variables associated with the $x_{ij.}$, $x_{i.k}$, and $x_{.jk}$ constraints, respectively. Some analogies between log-linear modeling and entropy maximization were already shown in the previous section (see also Willekens, 1980a). The dual variables of the entropy maximization may be expressed in terms of the u -parameters of the log-linear model and vice versa. The multiproportional adjustment algorithm is slightly

different if another combination of marginal totals is given. For instance, consider the 1FE-case, with $x_{ij.}$ and $x_{..k}$ given, the algorithm then goes as follows:

Step 0: $s = 0$

$$m_{ijk}^{(0)} = m_{ijk}^0 \quad \text{for all } i, j, k \in S.$$

Step 1: For each k -value, adjust the (i, j) -matrix to the total $x_{..k}$

$$\hat{m}_{ijk}^{(2s+1)} = \hat{m}_{ijk}^{(2s)} \frac{x_{..k}}{\sum_{i,j} \hat{m}_{ijk}^{(2s)}} \quad \text{for all } k.$$

Step 2: Adjust along the age dimension k

$$\hat{m}_{ijk}^{(2s+2)} = \hat{m}_{ijk}^{(2s+1)} \frac{x_{ij.}}{\sum_k \hat{m}_{ijk}^{(2s+1)}} \quad \text{for all } i \text{ and } j.$$

Unless the stopping criterion is reached, $S = s + 1$ and go to step 1. The estimates are MLEs with main effects and the (AB)-interaction effect derived from the marginal totals and with the (AC), (BC), and (ABC) effects "borrowed" from the preliminary estimates, i.e., $u_{ik}^{AC} = {}^0u_{ik}^{AC}$; $u_{jk}^{BC} = {}^0u_{jk}^{BC}$ and $u_{ijk}^{ABC} = {}^0u_{ijk}^{ABC}$. The log-linear model (34) reduces to

$$\ln(m_{ijk}/m_{ijk}^0) = r_u + r_{u_i}^A + r_{u_j}^B + r_{u_k}^C + r_{u_{ij}}^{AB}$$

or

$$m_{ijk} = m_{ijk}^0 r_w r_{w_i}^A r_{w_j}^B r_{w_k}^C r_{w_{ij}}^{AB}.$$

Analogously, an algorithm can be derived for the 3E-problem.

This problem was studied by Evans and Kirby (1974) in an attempt to generalize spatial interaction (gravity) models developed in transportation science in order to infer area-to-area traffic flows by types of goods. The Evans-Kirby model is as follows:

$$m_{ijk} = r_i s_j p_k m_{ijk}^0,$$

with r_i , s_j , and p_k balancing factors. This model is nothing else but the log-linear model of the 3E-case, with only the main effects present in the x-terms.

Note that any set of preliminary estimates that exhibits the same interaction effects not found in the x-terms (marginals) leads to the same final estimates (for a formal proof, see Bishop *et al.*, 1975, p. 93). Hence, in the 3F-case, any array with given u_{ijk}^{ABC} -values yields the same estimates. This is not surprising since the only contribution of the initial array to the final estimates is the second-order interaction.

The multiproportional adjustment algorithm has a long history. In the statistical literature, it is often associated with Deming and Stephan (1940) and referred to as the "classical" iterative proportional fitting procedure (Bishop *et al.*, 1975, p. 84). For two-dimensional tables (matrices), however, the method was also independently developed in other areas of scientific inquiry to solve estimation problems. According to Murchland (1978), the earliest application of this technique of biproportional fitting was in forecasting telephone traffic to evaluate the needs for network expansion and dates back to 1937 (Kruithof, 1937). The algorithm is widely used in transportation science where it was introduced in 1954 by Fratar in

an attempt to forecast area-to-area person transportation demand. Since then, transportation science, where the technique is also known as Furness procedure, elaborated the method in order to estimate spatial interaction flows. The direction of elaboration was a reformulation as a gravity model (see below). Transportation scientists also applied the gravity model to infer migration flows and the technique was later adopted by migration analysts. A further independent development of the biproportional adjustment procedure is due to Leontief (1940) and Stone (1962), both in the economic field. In input-output modeling, the technique is known as RAS method and associated with Stone, and as nonsurvey technique for estimating input-output tables. It has been investigated extensively (Bacharach, 1970; MacGill, 1977). The link with the log-linear model of categorical data has however not been made. As a consequence, the contribution of each source of prior information to the estimates could not be unraveled.

In the RAS technique, the model format is as follows (for a two-dimensional table):

$$m_{ij} = r_i s_j m_{ij}^0 . \quad (37)$$

The balancing factors r_i and s_j must be determined from the data. Because the sum of row totals must be equal to the sum of column totals, the balancing factors are unique up to a scalar value. Various scaling procedures may be imagined. In the illustrations in Section 5, we postulate $s_1 = 1$. Stone (1962) suggested the following iterative procedure:

Step 0: $s(\text{step}) = 0$

$$r_i^{(0)} = 1 .$$

$$\text{Step 1: } s_j^{(2s+1)} = \frac{x_{.j}}{\sum_i r_i^{(2s)} m_{ij}^0} \quad \text{for } j = 2, 3, \dots, C .$$

$$\text{Step 2: } r_i^{(2s+2)} = \frac{x_i}{\sum_j s_j^{(2s+1)} m_{ij}^0} \quad \text{for } i = 1, 2, \dots, R .$$

Unless the stopping criterion is reached, $s = s + 1$ and go to step 1. The Frater or Furness technique follows the same algorithm. As was already pointed out, the bi-(multi-)proportional adjustment problem may also be solved by maximizing an entropy function. In the two-dimensional case, the entropy solution is

$$m_{ij} = m_{ij}^0 \exp[-(1 + \lambda_i + \mu_j)] , \quad (38)$$

where λ_i and μ_j are dual variables associated with the row constraints and column constraints, respectively.³

³Sometimes, (38) is written as

$$m_{ij} = m_{ij}^0 \exp[\hat{\lambda}_i + \hat{\mu}_j] .$$

Both expressions are equivalent since one may define $\hat{\lambda}_i = 1 - \lambda_i$ and $\hat{\mu}_j = -\mu_j$. Whereas (38) follows from maximizing the entropy

$$W = \sum_{i,j} m_{ij} \ln(m_{ij}/m_{ij}^0) ,$$

the above expression follows from minimizing the negentropy ($-W$), also referred to as information gain or information divergence. The structural similarity between biproportional adjustment and information minimization is therefore also demonstrated.

The balancing factors of the RAS-technique and the dual variables or Lagrange multipliers of the entropy method are related to one another. We may write:

$$\lambda_i = -(1 + \ln r_i), \quad \mu_j = -\ln s_j.$$

By relating the multipliers and the balancing factors to the log-linear model parameters, they may be given a statistical interpretation. Since

$$-[1 + \lambda_i + \mu_j] = r_u + r_{u_i}^A + r_{u_j}^B,$$

and since $\sum_j r_{u_j}^B = 0$, we have

$$\sum_j -[1 + \lambda_i + \mu_j] = C r_u + C r_{u_i}^A,$$

and, consequently,

$$\begin{aligned} \lambda_i &= - \left[1 + \frac{1}{C} \sum_j \mu_j + r_u + r_{u_i}^A \right], \\ \mu_j &= - \left[1 + \frac{1}{R} \sum_i \lambda_i + r_u + r_{u_j}^B \right], \end{aligned} \quad (39)$$

with C and R the number of, respectively, columns and rows in the matrix. Expressing the log-linear model parameters in terms of the Lagrange multipliers gives:

$$\begin{aligned} r_u &= - \left[1 + \frac{1}{R} \sum_i \lambda_i + \frac{1}{C} \sum_j \mu_j \right], \\ r_{u_i}^A &= - \left[1 + \lambda_i + \frac{1}{C} \sum_j \mu_j \right] - r_u \\ r_{u_j}^B &= - \left[1 + \frac{1}{R} \sum_i \lambda_i + \mu_j \right] - r_u. \end{aligned} \quad (40)$$

The equalities hold for any scaling applied to the balancing factors and to the Lagrange multipliers.

4.2.2. *Obtaining Cell Estimates from Marginal Totals*

When Some Cell Values Are Fixed

In the previous section it was assumed that none of the individual x_{ijk} -values was known exactly. In practice, it happens frequently that some cell counts are given. The data set used by Schoen and Woodrow (1980) is a case in point.

Recall that S is the set of (i,j,k) -cells that are not fixed. The treatment of $(i,j,k) \notin S$ is very simple. An $\{m_{ijk}^0\}$ -array is constructed that contains a zero value in all (i,j,k) -cells not in S . To construct $\{m_{ijk}^0\}$, first all structural zeros ($x_{ijk} = 0$) are entered into the account of preliminary estimates. Next, non-zero x_{ijk} 's are subtracted from the associated marginal totals and then removed from the table; a zero is entered in the appropriate cells of the $\{m_{ijk}^0\}$ -account. The remaining cells are given a value of one (or any value expressing a desired interaction effect). This procedure yields the following accounts:

$$(a) \quad m_{ijk}^0 = 0 \text{ for cells } (i,j,k) \notin S,$$

$$m_{ijk}^0 = 1 \text{ for cells } (i,j,k) \in S,$$

(b) the $\{x_{ijk}\}$ -account contains the revised marginal totals.

The expected m_{ijk} -values are computed by the multiproportional adjustment algorithm. However, only cells in S are taken into account and hence structural zeros are preserved in the estimation procedure. The interaction structure imposed on the estimates is one of *quasi-independence* (Bishop et al., 1975,

p. 179). Quasi-independence implies independence between the cross-classified variables provided that we do not consider rows, columns, or layers that have structural zero entries in at least one of the cells. The log-linear model of a table containing structural zeros is defined for the subtable with cells $(i,j,k) \in S$; the parameters are computed on the basis of the nonzero cells only.

This simple procedure may pave the way for a new combination of observed and estimated data. Cell values for critical categories may be obtained from a special survey, whereas the remaining cell values can be estimated by other means. Hewings and Janson (1980) propose such a procedure for input-output analysis.

4.2.3. *Constructing Accounts of Preliminary Estimates*

Patterns of interaction between cross-classified variables, which are not contained in the x-terms, may be imposed on the final estimates through the account of preliminary estimates. Any $\{m_{ijk}^0\}$ -account that exhibits the desired interaction patterns yields the same set of final estimates. In estimation problems, the true interaction patterns, i.e., those of the $\{x_{ijk}\}$ -account, are not all known. In general, higher-order interaction patterns are unknown and can only be approximated by choosing an appropriate $\{m_{ijk}^0\}$ -array. Several strategies may be adopted, only a few are mentioned here.

A. *Outdated table.* If structural change is minor, then the $\{m_{ijk}^0\}$ -account may consist of a table collected at a prior data. The table as such may be outdated, but the interaction pattern may still be valid. By adjusting the old table to new marginals, the higher-order interaction pattern is kept. This

procedure is generally followed in input-output analysis, where the preliminary estimates are generated by a survey held at a prior date, and the new marginals are taken from the national accounts. The same procedure is proposed by Shulman and Chaddha (1978) and Shulman (1979) to infer detailed population characteristics in an intercensal period from a combination of up-to-date aggregate data and outdated but detailed census-tabulations. In multidimensional population analysis, the data frequently come from censuses or special surveys held periodically but with intervals of considerable length. Adequate data sets may be generated by updating census information using recent aggregate data.

B. Table of intermediate/explanatory variables. Another approach is to derive a pattern of interaction from a cross classification of intermediate or explanatory variables. In this case, the account of preliminary estimates consists of intermediate or explanatory variables, and the higher-order interaction effects in this account are assumed to apply to the account of final estimates. This approach is widely used in spatial interaction analysis, although only implicitly. The spatial interaction models, developed in transportation science and regional science, have never been looked at as discrete multivariate analysis techniques. Only very recently were the analogies with models of categorical data realized (Willekens, 1980a). The disaggregated choice models for spatial interaction analysis, which are becoming increasingly popular, are also more closely tied to categorical data analysis, in particular to logit and probit analysis (McFadden, 1978).

In spatial interaction modeling, the challenge is to obtain accurate estimates of area-to-area flows of goods (transportation, traffic) or people (migration) on the basis of some aggregate data and some information on measures of spatial "friction," "impedance," or "deterrence." To derive these estimates, analysts have generally used the gravity model. The origin of the gravity model is not in statistics but in mechanics.

In the late nineteenth century, scientists such as Carey and Ravenstein pronounced the analogy between physical phenomena governed by Newton's law of gravity and social interactions. In 1948, Stewart formalized the idea by proposing that the interaction between place i and place j is directly related to the number of people in both places and is inversely related to the squared distance:

$$m_{ij} = gP_iP_j/d_{ij}^2, \quad (41)$$

where g is a constant to be determined from the data. Several efforts have been devoted to modifying the social gravity model without endangering its fundamental structure. The gravity model is a log-linear-model and may be written in terms of the model discussed earlier:

$$m_{ij} = w w_i^A w_j^B m_{ij}^0, \quad (42)$$

with $w = g$, $w_i^A = P_i$, $w_j^B = P_j$, and $m_{ij}^0 = d_{ij}^{-2}$. This transformation clarifies the role each term plays in the gravity model. The population figures generate the main effects; the distance term contributes the first-order interaction effect to the

$\{m_{ij}\}$ -matrix. Distance is a spatial friction factor, inhibiting interaction. The modifications of the gravity model, made in an attempt to increase the accuracy of the estimates, affect the way the main effects are measured and the first-order interaction is described. A review of the approach may be found in Wilson (1970) and, more recently, in Hua and Porell (1979). Particularly interesting to follow are the designs of spatial distribution functions $\{m_{ij}^0\}$, normally represented as F_{ij} . March (1971) summarizes the main types of functions proposed in the literature. Some of the functions commonly used are

(i) Inverse power function

$$F_{ij} = d_{ij}^{-\gamma} \quad (43)$$

(ii) Negative exponential function

$$F_{ij} = \exp[-\delta d_{ij}] \quad (44)$$

(iii) Tanner function

$$F_{ij} = \exp[-\delta_1 d_{ij}] d_{ij}^{-\delta_2}, \quad (45)$$

where γ , δ , δ_1 , and δ_2 are parameters to be estimated [Batty and Mackie (1972) provide an extensive overview of estimation procedures]. Instead of distance d_{ij} , transportation cost from i to j , c_{ij} , may be used. Whatever distribution function is used, its main purpose remains the same, namely, to get the best quantitative approximation of the higher-order interaction patterns, assumed to be present in the account to be estimated. The distribution function affects the estimates in a way which is completely analogous to the account of preliminary estimates.

As a result, research finding gathered in spatial interaction analysis may fruitfully be applied to categorical data analysis and vice versa.

5. NUMERICAL APPLICATIONS

To illustrate the techniques presented in this chapter, they are applied to two sets of mobility data. The first set consists of age-aggregated social mobility data of Britain and Denmark (Table 3). This simple application allows us to pay attention to the log-linear model parameters and their relationship to the balancing factors of the biproportional adjustment method and to the dual variables of the entropy maximization problem without getting lost in the sheer volume of numbers. The data set has already been thoroughly investigated by Bishop *et al.* (1975) from the perspective of testing for the presence of a particular pattern of interaction in the data.⁴ In this chapter, the interest is in the estimation of missing elements. Comparison of the worked examples of both perspectives illustrates a major point made in this chapter, namely, that estimation of missing elements in cross-classified data is equivalent to testing hypotheses on statistical independence. An additional consideration for selecting this illustration is that the study of social mobility patterns is an area to which multidimensional demographic techniques may fruitfully be applied, if an age dimension can be added to

⁴A similar analysis was carried out by Hauser (1979) for US-mobility data.

TABLE 3. Observed Danish and British Social Mobility Data^a

Father's status	Son's status					Total
	1	2	3	4	5	
A. Danish data						
1	18	17	16	4	2	57
2	24	105	109	59	21	318
3	23	84	289	217	95	708
4	8	49	175	348	198	778
5	6	8	69	201	246	530
Total	79	263	658	829	562	2391
B. British data						
1	50	45	8	18	8	129
2	28	174	84	154	55	495
3	11	78	110	223	96	518
4	14	150	185	714	447	1510
5	3	42	72	320	411	848
Total	106	489	459	1429	1017	3500

^aSource: Bishop et al., 1975, p. 100.

social mobility tables (the mobility figures in Table 3 are life-time measures). The second data set consists of age-specific migration flows of Austria and Sweden. Both countries publish age-specific flows by region of origin and of destination. The $\{x_{ijk}\}$ -account is therefore completely known. Willekens et al. (1979) used these data sets to test the validity of estimation methods they developed to infer the necessary migration data for multiregional population analysis. The main results are summarized here.

TABLE 4. Parameters of Saturated Log-Linear Model

Father's status	Son's status					Row effect
	1	2	3	4	5	
A. Danish social mobility table						
1	1.94	0.92	-0.07	-1.35	-1.46	-1.67
2	0.42	0.93	0.04	-0.47	-0.92	0.15
3	-0.33	0.00	0.31	0.13	-0.11	0.85
4	1.21	-0.36	-0.01	0.78	0.80	0.67
5	-0.82	-1.50	-0.27	0.90	1.69	0.00
Column effect	-1.17	-0.21	0.72	0.62	0.03	3.78
B. British social mobility table						
1	2.47	0.62	-0.83	-1.01	-1.24	-1.23
2	0.45	0.54	0.08	-0.31	-0.76	0.21
3	-0.38	-0.16	0.46	0.17	-0.09	0.10
4	-0.96	-0.33	0.15	0.51	0.62	0.92
5	-1.57	-0.67	0.14	0.64	1.47	0.00
Column effect	-1.51	0.23	-0.04	0.95	0.37	4.18

5.1. TWO-DIMENSIONAL SOCIAL MOBILITY TABLES

Table 4 shows the values of the parameters of the saturated log-linear model

$$\ln m_{ij} = u + u_i^A + u_j^B + u_{ij}^{AB},$$

where variable A is the father's status, variable B is the son's status. Note that the parameters of the multiplicative model can easily be derived from the u-parameters as shown in Table 1.

The parameter estimates are computed by the formula shown in Table 1, with the index set $K = \{1\}$, i.e., only a single layer is considered. A special computer program was developed,

although the estimates could also be obtained by applying ECTA and GLIM packages. The program also computes indicators not given by the standard programs and allows a flexible entry of prior information. The overall effect $u = 3.7831$ is shown in the lower right corner. The last column contains the u_i^A -parameters; they represent the effect of size differences in social status categories of fathers (row effects). The last row shows the u_j^B -terms, which measure the effects of size differences in son's status categories. The interaction terms u_{ij}^{AB} constitute the matrix elements. A negative u_{ij}^{AB} -term indicates that less sons are in status j , given that the father's status was i , than one would expect if son's status and father's status would be independent. The parameter values demonstrate that most sons remain in their father's status and that, if they change status, the change is generally to an adjacent status. Note that

$$\sum_i u_i^A = \sum_j u_j^B = \sum_i u_{ij}^{AB} = \sum_j u_{ij}^{AB} = 0,$$

as requested by the constraint specifications in Table 1. Tables 5 and 6 present estimates of cell counts under varying conditions of information availability. To demonstrate the impact of incomplete data on the interaction patterns exhibited by the estimates ($\{m_{ij}\}$ -account), the tables also give the log-linear model parameters associated with the various data availability conditions. In addition, the balancing factors of the adjustment model (RAS technique) adopted in the input-output analysis and the dual variables or Lagrange multipliers of the entropy method are shown.

TABLE 5. Danish Social Mobility Table Estimated from Marginal Totals Only ($m_{ij}^0 = 1$ for i, j)

Father's status	Son's Status					Total
	1	2	3	4	5	
A. Estimates						
1	1.9	6.3	15.7	19.8	13.4	57
2	10.5	35.0	87.5	110.3	74.7	318
3	23.4	77.9	194.8	245.5	166.4	708
4	25.7	85.6	214.1	269.7	182.9	778
5	17.5	58.3	145.9	183.8	124.6	530
Total	79	263	658	829	562	2391
						Row effect
B. Log-linear model parameters						
1	0.00	0.00	0.00	0.00	0.00	1.82
2	0.00	0.00	0.00	0.00	0.00	-0.10
3	0.00	0.00	0.00	0.00	0.00	.70
4	0.00	0.00	0.00	0.00	0.00	.80
5	0.00	0.00	0.00	0.00	0.00	0.41
Column effect	-1.53	-0.32	0.59	0.82	0.44	3.98
C. Balancing factors (r_i, s_j) and Lagrange multipliers (λ_i, μ_j)						
	r_i	λ_i	s_j	μ_j		
1	1.88	-1.63	1.00	0.00		
2	10.51	-3.35	3.33	-1.20		
3	23.39	-4.15	8.33	-2.12		
4	25.71	-4.25	10.49	-2.35		
5	17.51	-3.86	7.11	-1.96		

D. Goodness-of-fit

Pearson chi-square: 754 for ± 6 degrees of freedom.

Table 5 shows that if prior information is limited to marginal totals only, the interaction effect will be absent in the estimates. The expected counts may be expressed as follows:

$$m_{ij} = \exp \left[u + u_i^A + u_j^B \right] .$$

For example,

$$m_{32} = \exp [3.9765 + 0.7030 - 0.3244] = 77.9 .$$

The estimates may also be expressed in terms of the balancing factors: $m_{ij} = r_i s_j$. For example, $m_{32} = 23.39 \times 3.33 = 77.9$, and in terms of the Lagrange multipliers, $m_{ij} = \exp [-(1 + \lambda_i + \mu_j)]$, which gives $m_{32} = \exp [-(1 - 4.15 - 1.20)] = 77.9$. Table 5 illustrates how the quality of the estimates increases when a social mobility table of another country is added to the package of prior information. The chi-square value drops from 754 to 68. By comparing Tables 3 and 5, one observes that the British data contribute the first-order interaction parameters. The impact of the British data on overall, row and column effects are evaluated by comparing the parameter values of Tables 4 and 5. The expected counts may be expressed in the following terms:

Log-linear model parameters:

$$m_{ij} = m_{ij}^0 \exp \left[r_u + r_{u_i}^A + r_{u_j}^B \right] ,$$

$$\begin{aligned} \bullet \quad \text{ex.: } m_{32} &= 78 \exp [(3.8189 - 4.1842) + (0.7712 - 0.1012) \\ &\quad + (-0.0722 - 0.2300)] \\ \bullet \quad &= 78 \exp [-0.3653 + 0.6700 - 0.3022] \\ &= 78 \exp (0.0025) = 78.2 . \end{aligned}$$

Balancing factors:

$$m_{ij} = m_{ij}^0 r_i s_j ,$$

$$\text{ex.: } m_{32} = 78 \times 1.6302 \times 0.6150 = 78.2 .$$

Lagrange multipliers:

$$m_{ij} = m_{ij}^0 \exp[-(1+\lambda_i+\mu_j)]$$

$$\text{ex.: } m_{32} = 78 \exp[-(1-1.4887+0.4861)] = 78.2 .$$

5.2. THREE-DIMENSIONAL MIGRATION TABLES

The migration data required for multiregional population analysis consist of age-specific flow data by region of origin and region of destination. Few countries tabulate these data on a regular basis. In the context of the Comparative Migration and Settlement Study project at IIASA, which aimed at a comparative study of population distribution patterns in the 17 IIASA National Member Organization Countries applying multiregional demographic techniques, Willekens *et al.* (1979) developed a methodology to infer the necessary migration data from available information. The methodology was based on the principle of entropy maximization, but yields the same results as the perspective adopted in this paper. Assuming various combinations of marginal totals (x-terms), estimates of the detailed migration flows were obtained by the multiproportional adjustment algorithm and related procedures. Since no intraregional migration flows are considered in multiregional analysis, they were left out of the estimation problem. The array

of preliminary estimates was as follows:

$$m_{ijk}^0 = 1 \quad \text{for } i \neq j, \text{ for all } k.$$

$$m_{ijk}^0 = 0 \quad \text{for } i = j, \text{ for all } k.$$

The log-linear model implicit in the analyses is therefore one of quasi-independence.

To test the validity of the techniques, an error (goodness-of-fit) analysis was carried out for Austrian and Swedish data.⁵ The number of age groups was 18 in both data sets; the number of regions was 4 in Austria and 8 in Sweden. Table 7 summarizes the main results. The 3F-case yields estimates that are remarkably accurate. The Austrian estimates are given in Table 8. The estimates do not exhibit a second-order interaction effect (see the 3F-case in Table 2). The quality of the estimates may therefore be explained by the near absence of a second-order interaction effect in the observed data. This would mean that the pairwise interaction between origin (A) and destination (B) is the same in each age group (C); in other words, that the spatial mobility pattern is relatively independent of age. This is quite realistic. The error analysis of the Austrian 3F-case reveals that about half of the cells were estimated with less than 4% error and almost two-thirds of the migration volume has an estimation error of less than 4%. Table 9 gives the complete results of the error analysis. An important observation is that about 69% of the total absolute percentage error is due to the

⁵The method (3F-case) was actually applied to estimate missing migration data for Bulgaria (Philipov, 1978), The Netherlands (Drewe and Willekens, 1980), and Belgium (Willekens, 1977; Tan, 1980).

TABLE 7. Austrian and Swedish Migration Tables, Estimated from Various Combinations of Marginal Totals: Error Analysis

Case	Available data	Degrees of freedom ^a	Pearson chi-square		Average absolute % error	
			Austria	Sweden	Austria	Sweden
3E	$\{x_{i..}, x_{.j.}, x_{..k}\}$	$IJK - I - J - K + 2$	18590	-	31.09	34.58
1FE	$\{x_{ij.}, x_{..k}\}$	$(IJ-1)(K-1) - KI$	3662	-	16.24	15.26
2F	$\{x_{ij.}, x_{.jk}\}$	$(I-1)(J-1)K - KI$	2006	-	12.08	11.90
3F	$\{x_{ij.}, x_{i.k}, x_{.jk}\}$	$(I-1)(J-1)(K-1) - KI$	371	1262	4.27	6.32

^aThe number of d.f. associated with a fitted model is found by subtracting the number of independent parameters used in the model from the total number of cells to which the model is being fitted. For the quasi-independence model with no-zero entries in the marginal configurations, the number of d.f. is equal to the number of d.f. associated with the complete tabel, minus the number of structural zeros. For both Austria and Sweden, $K = 18$; for Austria $I = J = 4$; and for Sweden $I = J = 8$. Note: The 3E-case does not preserve the structural zeros.

TABLE 9. Error Analysis of (3F) Migration Estimates, Austria

A. Analysis by size class (flow volume) and migrant category

Size class	Number of flows ^a		Volume of flows		Cum. abs. % error		chi-square	
	total	-%-	total	-%-	value	-%-	value	-%-
0- 200	112.	51.85	8452.	10.63	1043.	68.56	0.9121e 02	33.70
200- 400	45.	20.83	12742.	16.02	241.	15.81	0.5711e 02	21.10
400- 600	20.	9.26	9481.	11.92	74.	4.87	0.2255e 02	8.33
600- 800	11.	5.09	7687.	9.67	73.	4.81	0.4191e 02	15.49
800-1000	9.	4.17	7705.	9.69	36.	2.36	0.1924e 02	7.11
1000-1200	3.	1.39	3330.	4.19	8.	0.52	0.2466e 01	0.91
1200-1400	7.	3.24	9075.	11.41	25.	1.63	0.1295e 02	4.78
1400-1600	1.	0.46	1464.	1.84	1.	0.06	0.1074e 00	0.04
1600-1800	0.	0.00	0.	0.00	0.	0.00	0.0000e 00	0.00
1800-2000	2.	0.93	3811.	4.79	7.	0.43	0.4201e 01	1.55
2000+	6.	2.78	15769.	19.83	14.	0.95	0.1887e 02	6.97
Total	216.	100.00	79516.	100.00	1522.	100.00	0.2706e 03	100.00

B. Analysis by error category

Category	Number of flows		Volume of flows		Average	
	error %	total	-%-	total	-%-	flow
1	0- 2	46.	21.30	24037.	30.23	522.543
2	2- 4	57.	26.39	24756.	31.13	434.316
3	4- 6	31.	14.35	13604.	17.11	438.839
4	6- 8	18.	8.33	6463.	8.13	359.056
5	8- 10	12.	5.56	4026.	5.06	335.500
6	10- 15	28.	12.96	5021.	6.31	179.321
7	15- 20	10.	4.63	798.	1.00	79.800
8	20- 30	10.	4.63	650.	0.82	65.000
9	30- 40	3.	1.39	158.	0.20	52.667
10	40- 60	1.	0.46	3.	0.00	3.000
11	60-100	0.	0.00	0.	0.00	0.000
12	100+	0.	0.00	0.	0.00	0.000
Total		216.	100.00	79516.	100.00	368.130

Average absolute percentage error = 4.27%
(relative mean deviation)

C. Analysis by size class and error category

Size class	0-2	2-4	4-6	6-8	8-10	10-15	15-20	20-30	30-40	40-60	60-100	100+	Total
0- 200	21	23	13	8	4	20	10	9	3	1	0	0	112
200- 400	9	12	10	5	3	5	0	1	0	0	0	0	45
400- 600	6	9	0	2	2	1	0	0	0	0	0	0	20
600- 800	1	2	4	0	2	2	0	0	0	0	0	0	11
800-1000	2	4	0	2	1	0	0	0	0	0	0	0	9
1000-1200	1	2	0	0	0	0	0	0	0	0	0	0	3
1200-1400	1	3	3	0	0	0	0	0	0	0	0	0	7
1400-1600	1	0	0	0	0	0	0	0	0	0	0	0	1
1600-1800	0	0	0	0	0	0	0	0	0	0	0	0	0
1900-2000	0	2	0	0	0	0	0	0	0	0	0	0	2
2000-	4	0	1	1	0	0	0	0	0	0	0	0	6
Total	46	57	31	18	12	28	10	10	3	1	0	0	216

^aThe structurally zero flows are excluded.

TABLE 8. Observed and (3F) Estimated Migration Flows by Age, Austria, Four Regions, 1966-1971

	Total	Migration from		East to		Total	Migration from		South to	
		East	South	North	West		East	South	North	West
0	1783.	0.	674.	874.	234.	0	882.	0.	556.	471.
		0-	-670-	-877-	-236-		-853-	0-	-575-	-481-
5	930.	0.	328.	482.	120.	5	493.	0.	348.	274.
		0-	-328-	-468-	-134-		-530-	0-	-329-	-256-
10	1597.	0.	483.	821.	293.	10	1371.	0.	1075.	1216.
		0-	-537-	-828-	-232-		-1342-	0-	-1044-	-1276-
15	4172.	0.	1351.	2029.	793.	15	3800.	0.	2022.	2501.
		0-	-1280-	-2192-	-700-		-3760-	0-	-1950-	-2613-
20	4227.	0.	1336.	2246.	645.	20	2097.	0.	1324.	1203
		0-	-1289-	-2231-	-707-		-2081-	0-	-1381-	-1163-
25	2807.	0.	939.	1477.	392.	25	1187.	0.	781.	656.
		0-	-910-	-1464-	-433-		-1225-	0-	-800-	-600-
30	1123.	0.	378.	596.	149.	30	465.	0.	333.	263.
		0-	-368-	-588-	-167-		-464-	0-	-346-	-252-
35	915.	0.	295.	493.	127.	35	392.	0.	281.	230.
		0-	-293-	-480-	-142-		-408-	0-	-276-	-219-
40	871	0.	280.	474.	117.	40	409.	0.	223-	174.
		0-	-312-	-438-	-121-		-411-	0-	-240-	-156-
45	579	0.	204.	306.	69.	45	277.	0.	140.	100.
		0-	-222-	-289-	-68-		-269-	0-	-149-	-99-
50	618.	0.	229.	314.	75.	50	256.	0.	147.	111.
		0-	-241-	-312-	-65-		-263-	0-	-134-	-117-
55	689	0.	261.	346.	81.	55	289.	0.	133.	99.
		0-	-280-	-331-	-78-		-296-	0-	-132-	-93-
60	700.	0.	259.	373.	67.	60	246.	0.	133.	76.
		0-	-269-	-357-	-74-		-253-	0-	-137-	-65-
65	543.	0.	203.	290.	51.	65	185.	0.	100.	55.
		0-	-210-	-280-	-53-		-190-	0-	-102-	-48-

70	353.	0.	131.	189.	33.	70	217.	118.	0.	64.	35.
		0-	-138-	-182-	-33-			-121-	0-	-65-	-31-
75	194.	0.	71.	105.	18.	75	118.	63.	0.	35.	19.
		0-	-74-	-101-	-19-			-65-	0-	-36-	-17-
80	68.	0.	24.	37.	6.	80	39.	22.	0.	11.	6.
		0-	-26-	-35-	-7-			-22-	0-	-12-	-5-
85	34.	0.	13.	18.	3.	85	21.	11.	0.	7.	3.
		0-	-13-	-18-	-3-			-11-	0-	-7-	-3-
Total	22203.	0.	7460.	11471.	3272.	Total	27773.	12564.	0.	7715.	7494.

	Total	Migration from		North to		Total	Migration from		South to	
		East	South	North	West		East	South	North	West
0	1445.	764.	402.	0.	280.	0	250.	352.	303.	0.
		-814-	-363-	0-	-268-		-229-	-395-	-281-	0-
5	917.	482.	252.	0.	184.	5	111.	155.	150.	0.
		-448-	-282-	0-	-187-		-108-	-124-	-184-	0-
10	1568.	745.	371.	0.	452.	10	148.	197.	222.	0.
		-771-	-344-	0-	-453-		-151-	-171-	-246-	0-
15	5297.	2888.	1107.	0.	1302.	15	736.	754.	750.	0.
		-2892-	-1123-	0-	-1282-		-772-	-809-	-659-	0-
20	3250.	1806.	734.	0.	710.	20	678.	736.	821.	0.
		-1861-	-701-	0-	-688-		-640-	-816-	-779-	0-
25	1885.	1029.	466.	0.	390.	25	395.	479.	499.	0.
		-998-	-482-	0-	-405-		-389-	-491-	-493-	0-
30	924.	492.	241.	0.	191.	30	165.	216.	226.	0.
		-485-	-255-	0-	-184-		-173-	-212-	-221-	0-
35	750.	402.	187.	0.	162.	35	113.	140.	155.	0.
		-379-	-213-	0-	-158-		-119-	-116-	-173-	0-
40	688.	419.	146.	0.	122.	40.	121.	113.	127.	0.
		-411-	-141-	0-	-136-		-128-	-87-	-146-	0-
45	447.	277.	101.	0.	68.	45	71.	69.	69.	0.
		-275-	-102-	0-	-70-		-81-	-50-	-77-	0-
50	478.	273.	124.	0.	81.	50	73.	88.	80.	0.
		-273-	-119-	0-	-86-		-65-	-81-	-95-	0-

	Total	Migration from		North to		Total	Migration from		South to	
		East	South	North	West		East	South	North	West
55	489.	304.	114.	0.	71.	237.	82.	82.	72.	0.
		-295-	-114-	0-	-80-		-84-	-64-	-89-	0-
60	439.	271.	110.	0.	58.	185.	59.	64.	61.	0.
		-263-	-114-	0-	-62-		-60-	-51-	-74-	0-
65	328	204.	83.	0.	41.	131.	42.	46.	43.	0.
		-198-	-84-	0-	-46-		-43-	-37-	-51-	0-
70	209.	130.	53.	0.	26.	82.	26.	29.	27.	0.
		-125-	-54-	0-	-30-		-28-	-21-	-33-	0-
75	109.	67.	27.	0.	14.	46.	15.	16.	16.	0.
		-66-	-27-	0-	-16-		-14-	-13-	-19-	0-
80	35.	22.	8.	0.	4.	14.	5.	5.	5.	0.
		-22-	-8-	0-	-5-		-5-	-3-	-6-	0-
85	19.	11.	6.	0.	2.	7.	2.	3.	2.	0.
		-11-	-6-	0-	-2-		-2-	-2-	-3-	0-
Total 19277.	10587.	4532.	0.	4158.	Total 10263.	3091.	3543.	3629.	0.	

Source: Observed migration data: Sauberer (1981); Estimates: Willekens, Por, and Raquillet (1979, pp. 34-35).

minor migration flows (less than 200 migrants), representing only 11% of the flow volume (Table 9a). A similar pattern is obtained if the chi-square statistic is used. The error distribution is, however, more explicit: the minor flows account for 34% of the total chi-square value. The contribution of minor flows to the overall error is further illustrated by the cross classification of error categories and flow size classes (Table 9c). Minor flows are concentrated in the large error categories. A similar observation was made by Bacharach (1970), Hewings (1977), Hinojosa (1978), and others in an error analysis of input-output coefficients, estimated by the RAS-method. The error analysis of minor flows raises two additional problems: the validity of the error measures used, and the effect of rounding flow estimates to the nearest integer value. Both problems received attention in the statistical literature. Suggestions for substituting the chi-square measure in case of small cell counts are many, as are suggestions for adjusting small cell values (less than 5, say). A pragmatic solution to the problems encountered may be to force small flows to be equal to the preliminary estimates. Its effect on the overall outcome would be negligible, and it would simplify the comparative evaluation of estimation methods. Hewings and Janson (1980, p. 847) propose such a strategy for the prediction of input-output tables.

The error analyses of the social mobility data and of the migration data show that an increase of the prior information leads to better estimates. However, each piece of prior information does not contribute equally. In the migration analysis, we could, for instance, observe that the estimates did not im-

prove substantially if we increased our knowledge of the age structure of immigrants (compare 2F-case with 1FE-case), whereas information on the overall migration pattern was essential (1 FE versus 3E). In the Danish social mobility analysis, knowledge of the British social mobility table had a very significant impact on the quality of the estimates.

Other authors have experienced similar observations. Snickars and Weibull (1977) compared the descriptive capacity of four alternative models of trip distribution among 12 regions in the Stockholm county. Each model used different amounts of prior information. An interesting observation was that a biproportional adjustment of a historical trip matrix to new marginals (Fratar method) outperformed the classical gravity model, with distribution function $\exp[-\beta c_{ij}]$ and c_{ij} the transportation cost. The application of a historical matrix instead of a cost function reduced the mean absolute percentage deviation from 20% to 7%, while the chi-square dropped from 731 to 107. The result indicates that trip patterns are not primarily shaped by travel cost differentials. The interaction effects exhibited by the distribution function are therefore inappropriate to describe the travel pattern. The conclusion of the authors that the gravity model has a smaller descriptive capacity than the Fratar model is, strictly speaking, incorrect. The outcome is not determined by the difference in model structure but by the difference in preliminary estimates. The application of the Fratar method with $m_{ij}^0 = \exp[-\beta c_{ij}]$ would give results identical to the gravity model. As these examples show, the value of each piece of prior information for estimation purposes is determined by the

relevance of the interaction effect it exhibits and imposes onto the estimates. Prior knowledge has no value in itself; it only contributes through the interaction effects it carries. The closer the patterns of association in the prior estimates resemble those in the data to be estimated, the better the estimates will be.

6. CONCLUSION

Multidimensional demography provides new opportunities for gaining demographic insight. The opportunities, however, can only fully be explored if statistical data are abundantly available or if appropriate estimation methods can be applied. Although the steps taken by statistical offices throughout the world to solve the data problem are encouraging, the lack of adequate data remains a major drawback of multidimensional population analysis. Estimation methods are therefore needed that are adapted to any particular data situation. This paper suggests a unified perspective on estimation techniques for multidimensional analysis with incomplete data.

A key feature of the unified approach is its emphasis on data structures and not on the values of individual data elements. The data set is approached as an interdependent, hierarchical system that can be modeled. The model relates the values, individual data elements take on, to structural characteristics of the data system; it is an aid in exploring data structures. In this perspective, missing elements are nothing but expressions of our incomplete knowledge of the systems

structure; and the problem of accurately estimating missing elements is a problem of hypothesizing the appropriate structure. To do this, optimal use should be made of all information available on the data system or on the phenomenon or process it represents. The techniques presented in this paper are intended to ease the formulation of structural hypotheses on the basis of the incomplete prior knowledge. An important advantage of the perspective and the techniques is that they are not only valid for conventional data sets (cross classifications of two or three variables), but that they can be applied equally well to multidimensional data sets.

To implement the unified perspective on estimating missing elements, an accounting framework is suggested. The advantages of accounts for multidimensional modeling and analysis were discussed by Rees (1980). Two types of accounts are distinguished. The first account contains what is known of the actual flow data. In general, prior knowledge is limited to marginal totals, structural zeros, and maybe a few elements. The second account contains preliminary estimates of the actual flows. The data in this account contribute to the quality of the estimates by imposing onto them patterns of association that are realistic but that cannot be derived from the limited information on the actual flows in consideration. How the mechanism works is easily seen from the log-linear models describing the two accounts (data sets). The introduction of an account of preliminary estimates m_{ijk}^0 is a compromise between maximum likelihood estimation under the independence model and maximum likelihood estimation assuming a saturated model. Through the $\{m_{ijk}^0\}$ -account, appropriate structural hypotheses with regard

to the estimates can be formulated. It is shown that these hypotheses may be derived from analogous data of a different but similar country, from historical data or from cross tabulations of intermediate or explanatory variables. The classical gravity model for estimating migration is an illustration of the latter case.

The estimation algorithm presented in this chapter is very simple, and it is not required to calculate the log-linear model parameters in order to determine the expected cell counts.

Values of the most probable flows, given the limited prior knowledge, are obtained by multiproportional adjustment of the preliminary estimates until they exactly fit the given information on the actual flows (first account or main data source). Multiproportional adjustment is not the only way to infer the required estimates. Other algorithms are discussed in the literature, but they are less transparent and therefore less appropriate if one wants to explore the consequences for the data structure and through it, for the individual estimates, if the available statistical, factual, and other knowledge of the phenomenon or system changes. This type of exploratory data analysis is however crucial for determining the contribution of each piece of prior information to the final estimates and for evaluating how much information is actually needed to ensure that a data set, required for the application of multidimensional demographic techniques, can be estimated with an acceptable level of accuracy.

ACKNOWLEDGMENTS

I very much wish to thank Willem Schaafsma and Paul de Jong for extensive comments on an earlier draft. The contribution of Jaine Koendering, who transformed the manuscript into a skillfully typed version, is greatly acknowledged.

REFERENCES

- Bacharach, M. (1970). *Biproportional Matrices and Input-Output Analysis*. London: Cambridge University Press.
- Batty, M. and Mackie, S. (1972). The calibration of gravity, entropy, and related models of spatial interaction. *Environment and Planning* 4, 205-233.
- Birch, M. (1963). Maximum-likelihood in three-way contingency tables. *Journal of the Royal Statistical Society B* 25, 220-233.
- Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: MIT Press.
- Caussinus, H. and Thelot, C. (1976). Note complémentaire sur l'analyse statistique des migrations. (Further note on the statistical analysis of migrations). *Annales de l'INSEE* 22-23, 135-146.
- Chilton, R. and Poet, R. (1973). An entropy maximizing approach to the recovery of detailed migration patterns from aggregate census data. *Environment and Planning A* 5, 135-146.
- Clogg, C. C. (1978). Adjustment of rates using multiplicative models. *Demography* 15, 523-539.
- Clogg, C. C. (1979). *Measuring Underemployment. Demographic Indicators for the United States*. New York: Academic Press.
- Deming, W. and Stephan, F. (1940). On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* 11, 427-444.
- Drewe, P. and Willekens, F. (1980). Maximum likelihood estimation of age-specific migration flows in the Netherlands. *Delft Progress Report* 5, 92-111.
- Evans, S. P. and Kirby, H. R. (1974). A three-dimensional Furness procedure for calibrating gravity models. *Transportation Research* 8, 105-122.
- Fienberg, S. E. and Mason, W. M. (1978). Identification and estimation of age-period-cohort models in the analysis of discrete archival data. In "Sociological Methodology 1979" (K. F. Schuessler, ed.), pp. 1-67. San Francisco: Jossey-Bass.
- Fisher, R. A. (1922). On the interpretation of Chi-square from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 85, 87-94.

- Fratrar, T. J. (1954). Forecasting distribution of interzonal vehicular trips by successive approximation. *Highway Research Board Proceedings* 33, 376-385.
- Gokhale, D. and Kullback, S. (1978). *The Information in Contingency Tables*. New York: Dekker.
- Goodman, L. (1978). *Analyzing Qualitative/Cathegorical Data*. Cambridge, Massachusetts, Abt Books, Abt Associates.
- Haberman, S. J. (1979). *Analysis of Qualitative Data (2 Vols)*. New York: Academic Press.
- Hauser, R. M. (1979). Some exploratory methods for modeling mobility tables and other cross-classified data. In "Sociological Methodology 1980" (K. F. Schuessler, ed.), pp. 413-458. San Francisco: Jossey-Bass.
- Hewings, G. J. D. (1977). Evaluating the possibilities for exchanging regional input-output coefficients. *Environment and Planning A* 9, 927-944.
- Hewings, G. J. D. and Janson, B. N. (1980). Exchanging regional input-output coefficients: a reply and further comments. *Environment and Planning A* 12, 843-854.
- Hinojosa, R. C. (1978). A performance test of the biproportional adjustment of input-output coefficients. *Environment and Planning A* 10, 1047-1052.
- Hoem, J. and Fong, M. (1976). A Markov chain model of working life tables. Working Paper no. 2, Laboratory of Actuarial Mathematics. Copenhagen University.
- Hua, C. and Porell, F. (1979). A critical review of the development of the gravity model. *International Regional Science Review* 4, 97-126.
- Jong, P. M. de (1981). The reliability of methods for predicting missing figures in migration tables. Paper prepared for presentation at the Conference on the "Analysis of Multidimensional Contingency Tables," Rome, June 25-26, 1981.
- Koesoebjono, S. (1981). Marital status life tables of female population in The Netherlands (1978); an application of the multidimensional demography. Working Paper no. 20, NIDI. Voorburg, The Netherlands.
- Kruithof, J. (1937). Calculation of telephone traffic. *De Ingenieur* 52, E 15-E 25. English translation by UK Post Office Research Department Library (No. 2663). London, Dollis Hill.
- Leontief, W. (1941). *The Structure of the American Economy, 1919-1939*. New York, Oxford University Press.

- In "Spatial Interaction Theory and Planning Models" (A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, eds.), pp. 75-96. Amsterdam: North-Holland.
- March, L. (1971). Urban systems: A generalised distribution function. In "Urban and Regional Planning" (A. G. Wilson, ed.). (London Papers in Regional Science, Vol. 2). London: Pion Ltd.
- Murchland, J. D. (1978). Application, History and Properties of Bi- and Multiproportional Models. London: University College, Traffic Studies Group, JDM-292.
- Nijkamp, P. (1979). Gravity and entropy models: the state of the art. In "New Developments in Modelling Travel Demand and Urban Systems. Some Results of Recent Dutch Research" (G. R. M. Jansen, P. H. L. Bovy, J. P. J. M. Van Est, and F. le Clerq, eds.), pp. 281-319. Westmead, Farnborough, England: Saxon House.
- Payne, C. (1977). The log-linear model of contingency tables. In "The Analysis of Survey Data. Vol. 2: Model Fitting" (C. O. Muirchear-
taigh and C. Payne, eds.), pp. 105-144. New York: Wiley.
- Philipov, D. (1978). Migration and settlement in Bulgaria. *Environment and Planning* 10, 593-617.
- Rees, P. H. (1980). Multistate demographic accounts: Measurement and estimation procedures. *Environment and Planning A* 12, 449-531.
- Rees, P. H. and Wilson, A. G. (1977). Spatial Population Analysis. London: Edward Arnold.
- Sauberer, M. (1981). Migration and settlement: Austria. Research Report, RR-81-16, IIASA, Laxenburg, Austria.
- Schoen, R. and Nelson, V. (1974). Marriage, divorce and mortality: A life table analysis. *Demography* 12, 313-324.

- annual meeting of the Population Association of America, Denver, Colorado, April 1980.
- Snickars, F. and Weibull, J. W. (1977). A minimum information principle. Theory and practice. *Regional Science and Urban Economics* 7, 137-168.
- Stein, R. L. (1980). National Commission recommends changes in labor force statistics. *Monthly Labor Review*, April 1980, 11-21.
- Stewart, J. Q. (1948). Demographic gravitation: Evidence and application. *Sociometry* 1, 31-58.
- Stone, R. (1962). Multiple classifications in social accounting. *Bulletin of the International Statistical Institute* 39, 215-233.
- Tan, E. (1980). On the estimation of migration flows by migrant categories. MA Thesis, Interuniversity Programme in Demography, Brussels.
- Willekens, F. (1977). The recovery of detailed migration patterns from aggregate data: An entropy-maximizing approach. Research Memorandum RM-77-58, IIASA, Laxenburg, Austria.
- Willekens, F. (1980a). Entropy, multiproportional adjustment and analysis of contingency tables. *Systemi Urbani* 2 (nr. 2-3), 171-201.
- Willekens, F. (1980b). Multistate analysis: Tables of working life. *Environment and Planning* 12, 563-588.
- Willekens, F., Por, A., and Raquillet, R. (1979). Entropy, multiproportional and quadratic techniques for inferring detailed migration patterns from aggregate data. Mathematical theories, algorithms, applications and computer programs. Working Paper WP-79-88, IIASA, Laxenburg, Austria (also published in *IIASA Reports* 4 (1981), 83-124.

Willekens, F., Shah, I., Shah, J. M., and Ramachandran, P. (1980).

Multistate analysis of marital status life tables. Theory and application. Working Paper no. 17, NIDI. Voorburg, The Netherlands.

(Revised version published in *Population Studies* 36 (1982), 129-144.

Wilson, A. G. (1970). Entropy in Urban and Regional Modelling. London: Pion Ltd.