WORKING WITH WHAT WE HAVE: HOW CAN EXISTING DATA BE USED IN THE CONSTRUCTION AND ANALYSIS OF SOCIO-DEMOGRAPHIC MATRICES?[1]

RICHARD STONE

University of Cambridge

This paper outlines a general strategy for constructing socio-demographic matrices, starting with a set of initial estimates based on available data and ending with a set of final estimates adjusted to meet the constraints connecting their true values.

The method is described and illustrated by a numerical example taken from the author's current work on marital transition matrices. The figures relate to the male population of England and Wales in 1978 and are based on British official statistics of population numbers, births, deaths, migrations, marriages, widowhoods and divorces.

1. INTRODUCTION

It is often said of the System of Social and Demographic Statistics (UNSO, 1975), and similar proposals for organizing socio-demographic data that they require an inordinate amount of statistical information, in particular information on human flows, and so cannot be implemented except where data have been specially collected for the purpose. I have never really believed this and when the SSDS Report was published my wife and I intended to carry out some experiments based on data for England and Wales. In 1975 we constructed and analysed a marital-status transition matrix relating to males in 1972. We finished the calculations but before they could be written up I had to put the work aside and in fact did not get back to it for five years. In 1980 I wrote a paper describing what we had done, not as a finished piece of work but as a basis for discussion with the statisticians in our Office of Population Censuses and Surveys on the availability of relevant official statistics and the use we had made of those which we knew about. These discussions have proved most helpful.

My intention had been to present to this conference a paper on marital transition matrices for a recent year, based not only on published material but also on such additional data as the OPCS could make available and taking into account a number of pitfalls, previously unknown to me, in the interpretation of various statistical returns. Things have not worked out as I intended for, while I have received and processed a great deal of information, I have not as yet compiled stock-flow matrices, let alone calculated any inverses. For that I am sorry, but from the viewpoint of this conference it may be no great loss: an account of the problems encountered in this area is probably of more general interest than a detailed set of results for a single year in one country.

[1]Paper presented to the 17th General Conference of the International Association for Research in Income and Wealth in August 1981.

2. The General Strategy

Working with what we have usually means that our data come from a number of sources intended originally to serve separate purposes and not requiring common definitions, classifications and conventions. So, what we have is likely to be in some measure inaccurate, inconsistent and incomplete. We are faced with much the same problem in constructing national accounts, as the existence of statistical discrepancies, residual errors, unidentified items and balancing entries bears witness. In either case we are dealing with a system and so we need a framework which shows how the parts of the system fit together and also the constraints to which they are subject. Given the framework, we can try to fill in the entries by direct estimation. Typically, we shall find that there are some gaps and at the same time some inconsistencies and we can then try to repair these deficiencies in the light of what we believe about the reliability of the various direct estimates. If we are successful in this endeavour we shall end up with a complete and coherent set of estimates which satisfy all the constraints to which the true values are subject. We are then ready for analysis.

Let us look at each of these aspects of the strategy in turn.

3. The Standard Stock-Flow Matrix

The framework described in this section is one I have used for many years and a version of it is set out in UNSO (1975, Table 7.1, p. 42). Here I shall expand it slightly to distinguish between those who enter and leave our country through birth and death and those who enter and leave it by immigration and emigration. In this presentation the standard stock-flow matrix appears as in Table 1.

TABLE 1

A Socio-Demographic Matrix Connecting the Opening and Closing Stocks of Year θ with the Flows During Year θ

State at New Year θ / State at New Year $\theta+1$	Other World	Rest of This World	Our Country: Opening States	Closing Stocks
Other World	α	δ	d'	
Rest of This World	β	γ	e'	
Our Country: Closing States	b	f	S	Λn
Opening Stocks			n'	

The symbols in this table have the following meaning:

α, a scalar, denotes the number of babies born in our country during year θ who die in our country before the end of it.

β, a scalar, denotes the number of babies born in our country during year θ who emigrate before the end of it.

b, a column vector, denotes the number of babies born in our country during year θ who survive in our country to the end of it.

Writing $i \equiv \{1\ 1\ 1 \ldots 1\}$ for the unit vector, then $\alpha + \beta + i'b$ denotes the total live births in our country during year θ.

δ, a scalar, denotes the number of immigrants into our country during year θ who die in our country before the end of it.

γ, a scalar, denotes the number of immigrants into our country during year θ who emigrate before the end of it.

f, a column vector, denotes the number of immigrants into our country during year θ who survive in our country to the end of it.

The sum $\delta + \gamma + i'f$ denotes the total immigrants into our country in year θ.

d', a row vector, denotes the deaths in our country in year θ of those who were present in it at the beginning of the year.

The sum $\alpha + \delta + d'i$ denotes the total deaths in our country in year θ.

e', a row vector, denotes the emigrants from our country in year θ who were present in it at the beginning of the year.

The sum $\beta + \gamma + e'i$ denotes the total emigrants from our country in year θ.

S, a square matrix, denotes the survivors in our country through year θ, and these are classified by their opening states in the columns and by their closing states in the rows.

n', a row vector, denotes the opening stock in each state. It can be seen that $n = d + e + S'i$ or, in other words, the people in our country at the beginning of the year either die there or emigrate in the course of the year or survive in it to the end of the year.

Λn, a column vector, denotes the closing stock in each state. The symbol Λ denotes the shift operator defined by the relationship $\Lambda^{\tau} n(\theta) \equiv n(\tau + \theta)$. It can be seen that $\Lambda n = b + f + Si$ or, in other words, the people in our country at the end of the year were either born in it or immigrated into it in the course of the year or were already in it at the beginning of the year and survived in it to the end.

The flows in Table 1 can be classified according to whether or not they form part of the opening and the closing stock. The people represented by α, β, γ and δ appear in neither; those represented by d and e appear in the opening but not in the closing stock; those represented by b and f appear in the closing but not in the opening stock; and those represented by S appear in both.

For those who prefer numbers to symbols, an example of Table 1 relating to the male population of England and Wales in 1978 is given in Table 2.

In describing the entries in Table 2 I shall begin with the opening stock of 23,980,900. Of these, 291,000 died and 93,000 emigrated, leaving 23,596,900 survivors in England and Wales at the end of the year. These survivors form the main component of the closing stock of 23,997,500 but to them must be added the new entrants of the year: 302,400 surviving births and 98,200 surviving immigrants.

The 302,400 births are not all the male births in England and Wales in 1978: to reach the total of 307,100 we must add in 3,800 newly born who died within the calendar year of their birth and 900 newly born who emigrated in the calendar year. Similarly, to reach the total of 99,300 immigrants we must add in 700 immigrant deaths and 400 re-emigrants in the calendar year of their immigration. Again, to reach the total of 295,500 deaths we must add in the

TABLE 2

A SOCIO-DEMOGRAPHIC MATRIX RELATING TO THE MALE POPULATION OF ENGLAND AND WALES IN 1978
(thousands)

State at New Year 1978 / State at New Year 1979	Other World	Rest of This World	England and Wales: Opening States	Totals
Other World	3.8	0.7	291.0	295.5
Rest of This World	0.9	0.4	93.0	94.3
England and Wales: Closing States	302.4	98.2	23596.9	23997.5
Totals	307.1	99.3	23980.9	

3,800 infant deaths and the 700 immigrant deaths, already referred to; and to reach the total of 94,300 emigrants we must add in the 900 infant emigrants and the 400 re-emigrants already referred to.

The framework provided by Table 1 can in principle accommodate any classifications of the population of our country by dividing the single row and column for our country into a number of rows and columns. If we are interested in marital status the indispensible criteria of classification are age and marital status. In the matrices on which I am at present working each sex is classified by single years of age up to age 84 and then to a single age group 85+. From age 15 onwards, five marital-status categories, single, first marriage, second or later marriage, widowed, divorced, are introduced into each age group. Thus in principle the transition matrices are of order 370 but in practice they are a little smaller because, in calculations to the nearest 100 persons, widowhoods, divorces and later marriages will not show up in the younger marriageable ages.

Having settled the question of a framework, there are a number of general taxonomic questions to which I shall now turn.

4. TAXONOMIC QUESTIONS

The questions I shall discuss in this section relate to the concepts of population and migration, the definition of age and some consequences of adopting age as a criterion of classification, and the difficulties raised by multiple transitions in a period.

(a) *The population.* The concept I shall adopt is usually called total population and consists of the people actually in the country plus members of the country's Forces serving overseas less members of other countries' Forces stationed in the country. This differs from the concept of normal residents by the difference between the number of foreign visitors and the number of normal residents on visits abroad.

(b) *Migrants and visitors.* Migrants are individuals whose displacement is intended to be permanent whereas visitors are individuals whose displacement is intended to be temporary. As far as possible the figures for immigrants and emigrants exclude visitors.

(c) *The definition of age.* Age is defined by reference to year of birth: anyone born in year θ is aged 0 at new year $\theta+1$ and aged 1 at new year $\theta+2$. Statistics of flows usually record age at the date the flow takes place, age at marriage, age at emigration, and age at death, and an adjustment is necessary to accord with the definition. This is usually straightforward but troublesome and could be avoided by using information on date of birth, which is usually asked for, to provide a retabulation according to the standard definition.

(d) *Age as a criterion of classification.* Although it is not essential for age to be a criterion of classification in socio-demographic accounts, it is a great advantage if it is, particularly if single years of age can be used. In the first place, with single years each column in the matrix relates to a single vintage (or cohort) and so no difficulties can arise where, in a changing population, transitions are a function of age. In the second place, certain changes of state may be prohibited before a certain age so that with grouped data only the last age in a group passes out of the prohibited zone. In all such cases transition probabilities will be wrongly estimated from age-grouped or age-free data unless adjustments are made.

(e) *Multiple transitions.* In a socio-demographic matrix drawn up for calendar year θ, states are only observed at new year θ and new year $\theta+1$, and so we may run into difficulties if an individual makes more than one change of state in the course of the year. For instance, if a man is divorced by his first wife and marries again within the year he should be shown as moving from first marriage to second marriage. We are unlikely to make this connection with what we have because marriage certificates do not contain information about the date of the preceding divorce. And we cannot discover anything by indirect estimation using the constraints because the implied changes in the entries cancel out so that the constraints can do nothing. It would seem, therefore, that in the absence of special information we are bound to work on the assumption that individuals make at most one change of state in a year.

Let us now turn to the first step in filling in the matrix, namely the use of primary data to form the direct estimates.

5. Filling in the Matrix: the Direct Estimates

The data available in recent years for constructing a marital-status transition matrix for males or females in England and Wales can be listed under the following headings: (a) opening and closing stocks; (b) live births; (c) deaths; (d) migrations; (e) marriages; (f) widowhoods; and (g) divorces. The methods used in making the direct estimates will now be briefly described.

(a) *Opening and closing stocks.* Estimates of total population classified by age and marital status as at 30 June are published, but estimates classified by age alone as at 31 December are also made by individual years of age up to age 94 and for the age group 95+. The mid-year information classified by age and marital status is available by single years of age up to age 84 and for the age group 85+. In order to make new-year estimates classified by age and marital status I propose to average, age by age, the mid-year marital status distributions and apply the average to the appropriate intermediate population component

for December 31. This will give me estimates by year of age for the ages 0–14, by year of age and marital status for the ages 15–84 and by marital status only for the age group 85+.

(b) *Live births.* Total live births occurring in a calendar year can be obtained from birth registrations. Babies surviving to the end of the calendar year of their birth can be estimated by deducting estimates, based on statistics of infant mortality, of babies who die in the calendar year of their birth and estimates, based on migration statistics, of babies who emigrate in the calendar year of their birth.

(c) *Deaths.* Registered deaths classified by year of age and marital status up to age 109 are available from death registrations, the normal time lag between occurrence and registration being a matter of days only. Age is defined as age at time of death and so a correction is needed to adjust the figures to the standard definition of age. A further, far less obvious, adjustment appears from survey data to be needed to allow for systematic tendencies to misreport marital status in death registrations. For example, there appears to be an excessive number of deaths classified to the married state with, in the case of males, deficiencies in all other states. Something can be done about this problem but there are others. For example, when a husband and wife die together, as might happen in a car accident, there is a legal convention that the husband dies before the wife. This may make good sense in legal terms but if carried into the statistics it implies that one married man and one widowed woman have died. The only way out of such an absurdity is to change the statistical convention.

(d) *Migrations.* Information on migrations into and out of England and Wales can be built up from several sources. The movements to and from all areas except Scotland and Northern and Southern Ireland can be obtained from the International Passenger Survey and some information is also available for movements to and from other parts of the British Isles. Data from these sources, adjusted to the standard definition of age, provide a basis for the required estimates.

(e) *Marriages.* Information on marriages is available for the first and second half of the calendar year classified by sex, age by single years up to age 84, and previous marital status. Thus first marriages can be distinguished from later marriages. Corrections are needed to convert the figures to the standard definition of age and to allow for presumed misreporting of marital status at marriage. The number of remarriages of divorced people is increased by a factor depending on age and sex, and the number of marriages of bachelors and spinsters is reduced correspondingly.

(f) *Widowhoods.* The number of men (women) widowed in a year is equal to number of deaths of married women (men). Information is available half-yearly by years of age up to age 84 and requires adjustment to the standard definition of age.

(g) *Divorces.* The information available on divorces is similar to that available on widowhoods and requires a similar adjustment for age.

This completes my account of the direct estimates. Let us now turn to the second step in filling in the matrix, namely the adjustment of the direct estimates and the estimation of entries for which no direct estimates are available.

6. Filling in the Matrix: the Indirect Estimates

Up to this point the matrix is incomplete for two reasons. First, the direct estimates, coming from different sources, may not be consistent; and, second, certain entries, namely the numbers remaining in the same marital state from one age to the next (that is throughout the year), have not been estimated at all. Both these defects can be remedied by adjusting the direct estimates to meet the constraints of the system by reference to a rating of reliability.

Before describing a formal solution of this problem it may be helpful if I give a numerical example of it in the present context. This is provided in Table 3 which sets out a part of the matrix relating to the ages 15 to 18.

TABLE 3

A Partial Matrix Relating to the Male Population of England and Wales in 1978 Illustrating Adjustment and Indirect Estimation

(thousands)

State at New Year 1978 / State at New Year 1979		Births	Immigrants	England and Wales: Opening States								Closing Stocks
				15	16		17		18			
				s	s	m_1	s	m_1	s	m_1		
Deaths		—	—	0.2	0.4	0.0	0.5	0.0	0.4	0.0		
Emigrants		—	—	2.5	2.5	0.0	2.5	0.0	2.5	0.5		
England and Wales: Closing States	15 s		1.5									416.4
	16 s		3.5	( )								408.6
	16 m_1		0.0	0.1								0.2
	17 s		4.0		( )							399.6
	17 m_1		0.0		0.9	( )						1.0
	18 s		4.5				( )					388.8
	18 m_1		0.0				4.5	( )				5.5
Opening Stocks				408.0	399.4	0.1	391.9	1.1	—	—		

The numbers in Table 3 relate to direct estimates of opening and closing stocks, deaths, immigrants and emigrants, and first marriages. At these ages there are only two marital states to be considered: single, s, and married for the first time, m_1. The five empty brackets indicate the entries for which indirect estimates are needed. The data show a high degree of coherence since it makes very little difference, if any, whether the indirect estimates are made from the row or the column in which they are located. I should mention, however, that all the entries for migration in this paper are only provisional.

The row for 16 m_1 is the only row (or column) in the table which, while complete, contains no empty bracket. It is subject to the constraint that the two flow entries should sum to the stock entry. But, as can be seen, this constraint

is not exactly met: $0.0+0.1 \neq 0.2$. Although this instance is trivial, in principle some adjustment is called for.

The other complete rows and columns all contain an empty bracket which could be filled in either of two ways. Denote by x_1 the entry at the intersection of column $15\,s$ and row $16\,s$. Then, from the column, $x_1 = 408.0-0.2-2.5-0.1=405.2$ and, from the row, $x_1 = 408.6-3.5=405.1$. Again the discrepancy is trivial but in principle some adjustment is called for. The constraint is obtained by equating the entries on the right-hand sides of the two equations, and the direct estimates of these entries can be adjusted to meet it by reference to their reliability ratings.

When we come to column $16\,s$ and row $17\,m_1$ we see a figure of 0.9 for the first marriages of males who were 16 years old at new year 1978. This item enters into the indirect estimation of two unknowns which we may call x_2 and x_3. We can eliminate each of these using the row and column equations into which they enter and we are left with two equations from which we can eliminate the 16-year-old marriages. By using the resulting equation to adjust the remaining direct estimates in all the equations, we can adjust the direct estimate of 16-year-old marriages and then make indirect estimates of the two items that were not directly estimated.

There is one further point to be made. In the course of carrying out the adjustments we can expect to alter the components of all the categories estimated directly. We might wish to impose further constraints to ensure that these components summed to given totals or did not depart from them by unduly large amounts in relation to their reliability.

Let us now turn to a formal statement of the adjustment procedure which can be used to handle all these problems.

7. The Adjustment Procedure and Its Formalization

In the preceding section the only form of constraint arose from the arithmetical identity that components sum to totals. Constraints might take other forms: estimates must be single-valued, accounts must balance, and so on.

It is clear that in adjusting the entries in a matrix we should not wish to change much those direct estimates which we believed to be relatively accurate whereas we should be willing to make considerable changes in those estimates which we believed to be relatively inaccurate. In order to carry out the adjustments, therefore, we should need reliability ratings of the direct estimates from which we could construct a variance matrix for them. We could then set out to minimize the sum of the squares of the adjustments, weighted by the reciprocals of the variances, which would enable the constraints to be met.

The most difficult part in all this is to construct a good variance matrix. Before coming to that question I shall first formalize the procedure I have outlined.

Let x, of type $\nu \times 1$, denote a vector of the true values of the unknowns which are subject to μ independent linear constraints given by

$$Gx = h \tag{1}$$

where G, the constraint matrix, is of type $\mu \times \nu$ and rank μ; and h, a vector of known constants, is of type $\mu \times 1$. Let x^* denote a vector of unbiased estimates of the elements of x; let V^*, of order ν and rank greater than μ, denote the variance matrix of the elements of x^*; and assume that any constraints satisfied by x^* are linearly independent of (1).

The best linear unbiased estimator, x^{**} say, of x can be written as

$$x^{**} = x^* - F(Gx^* - h) \tag{2}$$

where F denotes an arbitrary matrix of type $\nu \times \mu$. The estimator x^{**} will satisfy (1) provided that

$$Gx^{**} - h = 0 \tag{3}$$

that is, from (2), provided that

$$(I - GF)(Gx^* - h) = 0 \tag{4}$$

for all values of x^*, and this requires that

$$GF = I. \tag{5}$$

The variance matrix, V^{**}, of x^* is

$$V^{**} = (I - FG)V^*(I - FG)' \tag{6}$$

and to obtain estimates of the elements of x^{**} with least variance we must minimize the diagonal elements of (6) subject to (5). From this it follows that F^*, the estimator of F, must satisfy the relationship

$$-V^*G' + F^*GV^*G' - G'L = 0 \tag{7}$$

where L denotes a matrix, of order μ, of Lagrange multipliers. If we premultiply (7) by G we see that $GG'L = 0$ since $GF^* = I$. Hence $L = 0$ since GG' is nonsingular. Consequently

$$F^* = V^*G'(GV^*G')^{-1} \tag{8}$$

which can always be formed since GV^*G' is also nonsingular. From (2) and (8)

$$x^{**} = x^* - V^*G'(GV^*G')^{-1}(Gx^* - h) \tag{9}$$

from which we see that V^* need only be known up to a scalar multiplier which will cancel out. From (6) and (8)

$$V^{**} = V^* - V^*G'(GV^*G')^{-1}GV^* \tag{10}$$

This, I think, is the traditional way of setting out the problem and its solution but it is not the only way. As is pointed out in Byron (1978), (9) can be obtained by combining the first-order conditions for minimizing a constrained quadratic loss function. Thus, denoting the loss by ω, the function can be written, in the notation used above, as

$$\omega = \tfrac{1}{2}(x^{**} - x^*)'V^{*-1}(x^{**} - x^*) + l'(Gx^{**} - h) \tag{11}$$

where l denotes a vector of Lagrange multipliers. Writing l^* for the estimator

of l, the first-order conditions for a minimum of (11) are

$$l^* = (GV^*G')^{-1}(Gx^* - h) \tag{12}$$

and

$$x^{**} = x^* - V^*G'l^*. \tag{13}$$

By substituting for l^* from (12) into (13) we obtain (9).

The significance of this reformulation lies in the computational possibilities it opens up. Procedures based on the conjugate gradient algorithm can be used in minimizing the loss function and these turn out to be much more efficient than the traditional methods of solving (9) in terms both of time taken and storage capacity in the computer. Thus it becomes practicable to adjust very large matrices and the decisive difficulty in carrying out adjustments is removed.

Let us now turn to the question of constructing the variance matrix V^*.

8. The Construction of the Variance Matrix

As with the entries in the national accounts, so in the present instance it is virtually impossible to measure the variances of the direct estimates. But statisticians, with experience in a particular area, usually acquire impressions about the reliability of the statistics they handle and sometimes formalize these impressions in terms of a reliability rating. This means that they assign the supposed percentage errors in the various direct estimates to ranges, such as <3 percent, 3–10 percent, >10 percent. This form of reliability rating recognizes that, within different sources, variances are likely to be proportionate to the square of the size of the estimate but that different sources are likely to be accompanied by different factors of proportionality. Such ratings are provided in Britain for the main aggregates in the national accounts in UKCSO (1968). But it may be possible to go further. For example, in Britain income from employment, private consumers' expenditure and public consumers' expenditure are all given an A-rating, signifying that their margins of error are within the range ±3 percent. Public consumers' expenditure is based on accounting data and should be substantially accurate; the other two aggregates are built up from a variety of sources of varying reliability and it seems likely that they would come at a lower point in the A-range than public consumers' expenditure. In any event, in using the reliability ratings to form a variance matrix assumptions have to be made about the point in the range appropriate to the different estimates. It can always be assumed that estimates come at the mid-point of their range but it is clearly an advantage if some discrimination can be made. It is also an advantage if values can be given to the covariances which will arise where estimates are not independent.

The construction of a variance matrix of the direct estimates is the last step needed to construct the stock-flow matrix in its standard form. One further change is wanted for analytical purposes and in the following section I shall describe this and the calculations which follow.

9. The Final Form of the Matrix and the Analytical Calculations

At this point we have a complete stock-flow matrix the entries in which satisfy the constraints connecting their true values. The change required for analytical purposes is that the vector of emigrants be deleted and its components be subtracted from the corresponding components of the vectors of immigrants and of opening and closing stocks. In terms of Table 1, we must suppress e' and replace f, n' and Λn by $f^* \equiv f - e$, $n'^* \equiv n' - e'$ and $\Lambda n^* \equiv \Lambda n - e$.

The reason for this change is that when we come to form the transition matrix by dividing the elements in the columns of S by the corresponding element of $n^{*\prime}$, the coefficients will be of the correct size since they will relate to survivors in our country divided by the opening stock of those who will either die or remain in our country in the coming year. If we used the elements of n' as divisors, the coefficients would be too small for all opening states from which there was any emigration. Deaths would be augmented by emigrants, and the estimates we made of life expectancies would be biased downwards, relating in fact to the expectation of life in our country and omitting that part of the expectation lived elsewhere. By using the elements of $n^{*\prime}$ as divisors we restrict the population on which the estimates of life expectancies are based to that part which will spend the rest of its life in our country.

With this change we have, corresponding to the row for our country in Table 1, the equation

$$\Lambda n^* \equiv b + f^* + Si \tag{14}$$

and if we denote the matrix of transition coefficients by C, then

$$C = S\hat{n}^{*-1} \tag{15}$$

where $\hat{n}^{*-1}$ denotes a diagonal matrix of the reciprocals of the elements of n^*. If the population were in stationary equilibrium, so that $\Lambda n^* = n^*$, then by combining (14) and (15) we could write

$$\begin{aligned} n^* &= b + f^* + Cn^* \\ &= (I - C)^{-1}(b + f^*) \end{aligned} \tag{16}$$

where the matrix multiplier $(I - C)^{-1}$ transforms the net new entrants of the year into the numbers in different states in the total population just as in economic input–output analysis the corresponding matrix multiplier (the Leontief inverse) transforms final demands into total outputs.

Demographically speaking the population of England and Wales was nearly stationary in 1978 and so the assumption that $\Lambda n^* = n^*$ which enabled us to derive (16) is nearly true. But for present purposes the stationarity or otherwise of the population is irrelevant because each column of C relates to a single age and so the elements of C are not affected by the age composition of the population as, in general, they would be if states were defined without reference to age.

If, further, we can assume that the elements of C are probabilities, that they are the same for everyone in a given state, then $(I - C)^{-1}$ can be interpreted as the fundamental matrix of an absorbing Markov chain.

The elements of $(I-C)^{-1}$ are times, but defined on the unrealistic assumption that deaths all take place at the end of the interval, in our case the end of 1978. It would be more realistic to assume that, on average, deaths take place half way through the interval. To make the necessary adjustments let us define a matrix $(I-C^*)^{-1}$ as

$$(I-C^*)^{-1} = \tfrac{1}{2}\tau(I+\hat{c})(I-C)^{-1} \tag{17}$$

where τ denotes the length of the interval expressed in the unit in which time is measured and so, in the present case, 1; and c denotes a vector of state-specific survival rates.

The column sums of $(I-C^*)^{-1}$ measure the expectation of life of an individual entering the state represented by the column, whereas the column sums of $(I-C)^{-1}$ exceed this expectation by half an interval. The elements of a column measure the expected time to be spent in different states by an individual entering the state represented by the column. Thus we can work out the expectation at birth of so many years in each of the marital states and we can repeat this calculation for an individual in any other state of the system.

The introduction of the inverse $(I-C^*)^{-1}$ invites us to return to the adjustment problem since it enables us to put what I have called indirect constraints on the elements of the survival matrix S. First, if in any column of $(I-C^*)^{-1}$ we add up the elements relating to different marital states at any particular age, the sum should not exceed 1 since this is the maximum number of years that can be spent in a single year of age. Inspection of the inverse shows at once any violations of this inequality constraint and these could be removed by the use of programming methods. Second, as I have said, the column sums of $(I-C^*)^{-1}$ measure life expectancies and so there would be constraints on these sums if we had independent measures of these expectancies.

An example of adjusting a socio-demographic matrix using both direct and indirect constraints is given in Stone (1975). In that example it turned out that the first type of indirect constraint was not violated and that the second type could be transformed into direct linear constraints on the entries in the stock-flow matrix. Thus all constraints could be imposed in a single operation. But, generally speaking, a more complicated, iterative solution is likely to be needed.

10. Conclusions

This is a purely methodological paper and so can only contain methodological conclusions. I have several to offer.

First, it seems likely that with the data available in England and Wales it will prove possible to construct fairly reliable marital-status transition matrices. It is true that some of the entries in these matrices are not estimated directly but they can be estimated indirectly by an application of the well-known method of adjusting conditioned observations. Although the method is over one hundred and fifty years old, computing methods have recently been improved and I have already referred to the paper by Byron (1978) on the subject.

Second, as things stand, the data available require a considerable amount of further processing. This is no criticism of official statistics since they have

been prepared for purposes different from mine. However, should the time come when it was desired to produce a general demographic framework for the construction of socio-demographic matrices, it would not be difficult to process and adjust the purely demographic data on stocks at new year, births, deaths, migrations and survivors, all classified by age and sex, and to incorporate this information in a continuing data bank. With this facility it would not be necessary to go into the basic demography every time a new socio-demographic matrix was to be constructed.

Third, in this paper the adjustment procedure is mainly used to estimate items for which there are no direct estimates. In principle it would be possible to make direct estimates: we could find out by survey methods the proportion of men or women aged η who remained throughout the year in a given marital status category. If we did this we should be nearer the position that has been reached in the national accounts, where most items are estimated directly. The adjustment method would then be used mainly to get rid of statistical discrepancies, that is it would return to its traditional function. This function seems to me important in the systematic use of national accounts data in model building. Economic parameters will be better estimated if we start from a consistent (and usually more accurate) set of accounts. The short-cut methods of achieving the appearance of consistency, such as adding the residual error to income while leaving expenditure unchanged or treating the unidentified items in the financial accounts as net acquisitions of financial assets, are likely to do more harm than good since they imply obvious misspecifications of error. The adjustment procedure described in this paper provides the basis for a more realistic apportionment of error, as can be seen from the preliminary attempt to adjust a version of the British national accounts in Stone (1982).

Finally, I should like to emphasise that, while the adjustment procedure is in principle clear and simple, in practice many questions arise on the best way to apply it because of the difficulties of stating exactly what we think we know and do not know about the direct estimates.

References

Byron, Ray P. The Estimation of Large Social Account Matrices, *Journal of the Royal Statistical Society, Series A*, vol. 141, pt. 3, pp. 359–367, 1978.

Stone, Richard, Direct and Indirect Constraints in the Adjustment of Observations. In *Nasjonalregnskap, Modeller og Analyse* (essays in honour of Odd Aukrust). Statistisk Sentralbyrå, Oslo, 1975.

———, Balancing the National Accounts. Paper presented to the Ivor Pearce Conference, Southampton, 1982. To be published.

U.K., Central Statistical Office (1968). *National Accounts Statistics: Sources and Methods*, H.M.S.O., London, 1968.

U.N., Statistical Office (1975). *Towards a System of Social and Demographic Statistics*, Studies in Methods, Series F, No. 18. U.N., New York, 1975.