Errors of Prediction for Markov Chain Models

By D. J. Bartholomew

London School of Economics and Political Science

[Received July 1974. Final revision March 1975]

Summary

When Markov chain models are used for predicting the state distribution the forecasts are liable to error for a variety of reasons. The sources of such error are identified and methods are given for finding the contribution which each source makes to the total error of prediction. In particular it is shown that the error arising from the estimation of the parameters is likely to be of the same order of magnitude as the inherent variation arising from the stochastic assumptions of the model. The theory is first derived for a closed system and then generalized to two kinds of open system.

Keywords: PREDICTION ERROR; MARKOV CHAINS; MANPOWER MODELS; DIRICHLET DISTRIBUTION; DIRECT MATRIX PRODUCT; FORECASTING

1. Introduction

The theory of Markov chains has found many applications in the social and management sciences for modelling processes of change. Examples include social and occupational mobility, voting behaviour, purchasing behaviour and educational and manpower planning. An account of much of this work will be found, for example, in Coleman (1964) and Bartholomew (1973). The matrix methods used for population projection are essentially deterministic versions of Markov theory and these have been in use for many years. More recently demographers have begun to consider the stochastic aspects of population processes and an account of some of these developments will be found in Pollard (1973).

The characteristic feature of all these problems is that individuals may be in one of a particular set of "states" at any time and that changes of state take place in a random manner. The central assumption of the discrete time Markov chain model is that individual movements are independent and are governed by transition probabilities which depend only on the current state and on the destination state. Even when this assumption is not justified for the problem as originally formulated it is often possible to re-define the states in such a way as to make it more nearly satisfied.

Markov models may be closed, meaning that no individuals enter or leave the system, or they may be open having both entry and loss. There may also be restraints on the total size of the system or on the input.

One of the main uses of the theory is for predicting the future numbers in the states—the stocks—under the assumption of constant transition rates. Such predictions are subject to error arising from several sources and it is with the nature of this error and its effect on the predictions that this paper is concerned. Our object is to identify the kinds of error, to provide formulae for calculating the variances and covariances of prediction error and to draw conclusions about the magnitude, relative and absolute, of errors arising from different sources.

2. PRELIMINARIES

Suppose that we have a closed system of N people distributed among k categories. Let the vector of the numbers in the categories at time T be

$$\mathbf{n}(T) = \{n_1(T), n_2(T), \ldots, n_k(T)\}.$$

Time will be treated as taking integer values for the purposes of accounting for the stocks but it is convenient to suppose that movements of individuals take place between these points. Thus $n_{ij}(T)$ will denote the number of individuals moving from state i to state j in the interval $(T, T+1)$. It immediately follows that

$$n_j(T+1) = \sum_{i=1}^{k} n_{ij}(T) \quad (j = 1, 2, \ldots, k), \tag{1}$$

where $n_{ii}(T)$ is interpreted as the number remaining in state i.

If the system is open an additional term must be added to (1) to accommodate the flow into state j from outside. For ease of exposition we shall concentrate first on the theory for closed systems making the extension to open systems in Section 8.

If p_{ij} is the probability of a movement from i to j between two adjacent time points then it is well known that

$$E[\mathbf{n}(T+1) \,|\, \mathbf{n}(T)] = \mathbf{n}(T)\,\mathbf{P}$$

and, hence, that

$$E\mathbf{n}(T) = \mathbf{n}(0)\,\mathbf{P}^{\mathrm{T}}. \tag{2}$$

This equation, and its generalizations, are the basis of most prediction exercises. The transition matrix $\mathbf{P}$ is usually estimated from current or historical data and $\mathbf{n}(0)$ represents the stocks in the base year. In practice, the actual stocks at time T will differ from the values predicted by the Markov model for a variety of reasons which we must now consider.

3. SOURCES OF ERROR

There are many sources of error in making forecasts using the Markov chain model and their relative importance will obviously depend on the field of application. A useful discussion of the matter in the context of population mathematics is given in Hoem (1973). He distinguishes three types of error which are further divided into six levels. Some aspects of Hoem's classification are specific to population projection but the main framework is equally relevant in other applications. For our present purposes a threefold classification, which has been used for some time in the manpower planning work of the British Civil Service, will be adequate. Its relationship to Hoem's will be indicated as we go.

First, there is the inherent randomness resulting from the stochastic assumptions of the model. Conditional on $n_i(T)$ the flow numbers from state i will be multinominal with probabilities $p_{i1}, p_{i2}, \ldots, p_{ik}$. Thus the stock numbers will be random variables with distributions determined by the multinomial character of the flows. As the projection is carried forward in time the error from this source will be compounded. This error we shall call the *random or statistical error* (Hoem's Level 2).

The theory for computing the moments and product moments of predictions subject to random error was first given by Pollard (1966) in the context of population mathematics; it was adapted for use in the social context of population mathematics by Bartholomew (1967, 1973). Applications of the theory in manpower planning have been given by Forbes (1971) and Sales (1971). Pollard's method, which forms the starting point of the development in this paper, depends on matrix methods with an elegant use of the direct matrix product (or Kronecker product). A different approach based on generating functions is used by Staff and Vagholkar (1971).

The second source of error arises from the fact that we rarely, if ever, know **P**. We therefore have to estimate **P** from past flow data. Predictions are then made using the equation

$$\hat{\mathbf{n}}(T) = \mathbf{n}(0)\,\hat{\mathbf{P}}^{T}, \tag{3}$$

where the circumflex on **P** indicates that its elements have been replaced by estimates; that on $\mathbf{n}(T)$ means that it is a prediction using estimated values of the parameters. This source of error is independent of the random error since it arises from past experience. This kind of error will be called *estimation error* (Hoem's Level 1).

The third kind of error, which we shall call *specification error* (Hoem's Levels 3–6), arises from the fact that the model may be wrongly specified in the sense that its assumptions may not hold in practice. There are many ways in which specification error may arise but here we shall confine attention to unpredictable changes in the parameters. When the transition rates are likely to change in a predictable way they can be expressed as functions of time and incorporated into the model. In other circumstances they may change in a more or less erratic way in response to the exigencies of the environment. It may then be more realistic to treat the elements of **P** as random variables taking different realized values in different time intervals (Hoem's Level 3).

Although estimation and specification error arise in quite different ways their mathematical treatment has much in common and, of the two, it is convenient to take specification error first.

4. Variances and Covariances of the State Numbers with a Random Transition Matrix

The analysis in this section starts from a straightforward extension of the theory of Bartholomew (1973, Chapter 2), also given by Pollard (1973, Chapter 9, equation (9.8.2)). Let $\boldsymbol{\mu}^*(T)$ be a $k(k+1)$-dimensional row vector having

$$En_1(T), En_2(T), \ldots, En_k(T)$$

in the first k positions and the expectations of the products $n_j(T)\,n_l(T)$ in the remaining k^2 positions. The latter are listed in dictionary order of their subscripts, that is

$$(1,1), (1,2), \ldots, (1,k), (2,1), (2,2), \ldots, (2,k), \ldots, (k,1), (k,2), \ldots, (k,k).$$

Then Pollard showed that

$$\boldsymbol{\mu}^*(T+1) = \boldsymbol{\mu}^*(T)\,E(\boldsymbol{\Pi}), \tag{3}$$

where $\boldsymbol{\Pi}$ is a matrix with dimension $k(k+1)\times k(k+1)$ and structure

$$\boldsymbol{\Pi} = \begin{array}{cc} \left[\begin{array}{c|c} \mathbf{P} & \mathbf{R} \\ \hline \mathbf{0} & \mathbf{P}\times\mathbf{P} \end{array}\right] & \begin{array}{c} \langle\cdot k\cdot\rangle \\ \langle\cdot k^2\cdot\rangle \end{array} \\ \begin{array}{cc} \langle\cdot k\cdot\rangle & \langle\cdot\cdot k^2\cdot\cdot\rangle \end{array} & \end{array} \tag{4}$$

The matrix $\mathbf{P}$ is the transition matrix, $\mathbf{R}$ is a $k \times k^2$ matrix with $\delta_{jl} p_{ij} - p_{ij} p_{il}$ in its ith row and (j, l)th column; $\mathbf{P} \times \mathbf{P}$ is the direct matrix product of $\mathbf{P}$ with itself having k^2 rows and k^2 columns with $p_{ij} p_{i'l}$ in the (i, i')th row and (j, l)th column. A useful reference on the direct matrix product and its properties is Graybill (1969, Chapter 8). The multiplication sign will be used to denote direct multiplication as in $\mathbf{P} \times \mathbf{P}$.

It is convenient to introduce a $k \times k^2$ dimensional matrix $\mathbf{J}$; the ith row of this matrix has a 1 in the (i, i)th column and zeros elsewhere. Premultiplication of a matrix with k^2 rows by $\mathbf{J}$ has the effect of deleting all rows labelled (i, i') for which $i \neq i'$. Note that $\mathbf{J}\mathbf{J}' = \mathbf{I}$. A simple calculation thus enables us to write $\mathbf{R}$ in (4) in the form

$$\mathbf{R} = \mathbf{P}\mathbf{J} - \mathbf{J}(\mathbf{P} \times \mathbf{P}). \tag{5}$$

The row sums of $\mathbf{R}$ are zero.

In order to facilitate comparison with the case of fixed $\mathbf{P}$, we shall express (3) in terms of a matrix $\mathbf{\Pi}_E$ having the same form as $\mathbf{\Pi}$ but with p_{ij} replaced everywhere by $E(p_{ij})$. This is the matrix we would use if we decided to ignore the variation in the p_{ij}'s and worked, instead, with the transition matrix of expected values. $E(\mathbf{\Pi})$ and $\mathbf{\Pi}_E$ are related as follows:

$$E(\mathbf{\Pi}) = \left[\begin{array}{c:c} E\mathbf{R} & \mathbf{R}_E - \mathbf{J}\mathbf{C} \\ \hdashline \mathbf{0} & (E\mathbf{P}) \times (E\mathbf{P}) + \mathbf{C} \end{array}\right]$$

$$= \mathbf{\Pi}_E + \begin{bmatrix} \mathbf{0} & -\mathbf{J}\mathbf{C} \\ \mathbf{0} & \mathbf{C} \end{bmatrix} = \mathbf{\Pi}_E + \mathbf{A}, \quad \text{say,} \tag{6}$$

where $\mathbf{R}_E$ is the matrix $\mathbf{R}$ with p_{ij} replaced by its expectation and $\mathbf{C}$ is a $k^2 \times k^2$ matrix having $\operatorname{cov}(p_{ij}, p_{i'l})$ in the (i, j)th row and (i', l)th column. If $\boldsymbol{\mu}(T)$ denotes the vector of first- and second-order central moments then after some straightforward algebra, we find that

$$\boldsymbol{\mu}(T+1) = \boldsymbol{\mu}(T)\,\mathbf{\Pi}_E + \boldsymbol{\mu}^*(T)\,\mathbf{V} + \boldsymbol{\mu}(T)\,(\mathbf{A} - \mathbf{V}), \tag{7}$$

where $\mathbf{V}$ is a matrix with structure

$$\mathbf{V} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}.$$

The effect of introducing uncertainty about $\mathbf{P}$ into the model is thus to add the last two terms of (7) to the difference equation. This equation shows that the second-order moments of the predicted stocks depend on the distribution of the p_{ij}'s only through their second-order moments. In principle, therefore, the problem is solved but in practice the translation of one's uncertainty about the elements of $\mathbf{P}$ into the large number of variances and covariances required for the calculation is a daunting prospect. Hoem (1972) reports that he and Schweder have made calculations of error for Norwegian population projections by estimating the variation of $\mathbf{P}$ from historical data. Our purpose is rather to gauge the relative magnitude of the errors, so far as possible, by mathematical analysis.

One possible simplification is to suppose that the rows of $\mathbf{P}$ are statistically independent. This means that the chances of moving out of any given state do not depend on variations occurring in flow rates from other states. In many applications this seems plausible, at least as a first approximation. Under this assumption the matrix **C** has zeros in the (i, i')th row unless $i = i'$. The assumption of independence greatly reduces the task of specifying the uncertainty about **P** but greater simplification in the analysis follows if we assume a particular form of distribution for each row of **P**.

5. Independent Dirichlet Distributions for the Rows of P

The Dirichlet distribution is commonly used for a set of proportions adding up to one and has been used in connection with Markov transition matrices by Martin (1967). According to this distribution, the density for the ith row has the form

$$f(p_{i1}, p_{i2}, \ldots, p_{ik}) = \frac{\Gamma(\alpha_{i1}+\alpha_{i2}+\ldots+\alpha_{ik})}{\Gamma(\alpha_{i1})\,\Gamma(\alpha_{i2})\ldots\,\Gamma(\alpha_{ik})} \prod_{j=1}^{k} p_{ij}^{\alpha_{ij}-1}$$

$$(\alpha_{ij} > 0 \text{ for all } j \text{ and } i = 1, 2, \ldots, k). \tag{8}$$

The main characteristics of the distribution can be deduced from its first- and second-order moments which are also all we require for the calculation of variances and covariances of the grade sizes. Thus

$$E(p_{ij}) = \alpha_{ij} \Big/ \sum_{h=1}^{k} \alpha_{ih}, \tag{9}$$

$$\operatorname{var}(p_{ij}) = E(p_{ij})\{1 - E(p_{ij})\}/D_i, \tag{10}$$

$$\operatorname{cov}(p_{ij}, p_{il}) = -E(p_{ij})\,E(p_{il})/D_i \quad (i, j = 1, 2, \ldots, k), \tag{11}$$

where $D_i = \sum_{h=1}^{k} \alpha_{ih} + 1$. The distribution is thus centred at a point determined by the relative values of the α's whereas the degree of concentration about that point is governed by the row totals of the α's.

Under the Dirichlet assumption the difference equation for $\boldsymbol{\mu}(T)$ simplifies considerably because of a relationship which then holds between the sub-matrix $\mathbf{R}_E$ of $\mathbf{\Pi}_E$ and **JC**. It follows directly from (10) and (11) that

$$\mathbf{R}_E = \mathbf{DJC}, \tag{12}$$

where **D** is a diagonal matrix with D_i as its (i, i)th element. Using this relationship and the fact that **C** and **J′JC** are identical when the rows of **P** are independent, **C** can be eliminated from (7). Omitting the first k elements of each vector and writing $\boldsymbol{\mu}_2(T)$ for the part of $\boldsymbol{\mu}(T)$ containing the second-order moments we have

$$\boldsymbol{\mu}_2(T+1) = \boldsymbol{\mu}_2(T)\,(E\mathbf{P} \times E\mathbf{P}) + \{E\mathbf{n}(T) - E\mathbf{n}(T)\,\mathbf{D}^{-1} + (E\mathbf{n}(T) \times E\mathbf{n}(T))\,\mathbf{J}'\mathbf{D}^{-1}\}\,\mathbf{R}_E. \tag{13}$$

The incorporation of Dirichlet variation for **P** into the model thus adds the two terms involving $\mathbf{D}^{-1}$ to (13).

6. Comparison of the Variance–Covariance Vectors for Fixed and Random P

If D_i is infinite for all i, (13) reduces to Pollard's result for the case of fixed **P**. The ith element of the k-dimensional vector which pre-multiplies $\mathbf{R}_E$ in (13) is easily shown to be $En_i(T)\{1 - D_i^{-1} + En_i(T)\,D_i^{-1}\}$, hence for the effect of variation in **P** to be negligible it is clear that D_i must be an order of magnitude greater than $En_i(T)$. If, on the other hand, D_i were equal to $En_i(T)$ the variances and covariances of the stock numbers would be roughly those for an organization double the size. We shall now attempt to make this conclusion more precise by considering the cases $T = 1$ and $T = \infty$ in more detail.

The Case $T = 1$

Since $\boldsymbol{\mu}_2(0) = \mathbf{0}$, (13) gives

$$\boldsymbol{\mu}_2(1) = \{E\mathbf{n}(0) - E\mathbf{n}(0)\,\mathbf{D}^{-1} + (E\mathbf{n}(0) \times E\mathbf{n}(0))\,\mathbf{J}'\mathbf{D}^{-1}\}\,\mathbf{R}_E \tag{14}$$

and hence

$$\left.\begin{aligned} \operatorname{var}\{n_j(1)\} &= \sum_{i=1}^{k} n_i(0)\left\{1 + \frac{n_i(0)-1}{D_i}\right\} Ep_{ij}(1 - Ep_{ij}), \\ \operatorname{cov}\{n_j(1), n_l(1)\} &= -\sum_{i=1}^{k} n_i(0)\left\{1 + \frac{n_i(0)-1}{D_i}\right\} Ep_{ij}\,Ep_{il}. \end{aligned}\right\} \tag{15}$$

In this case, having $D_i = n_i(0)$ for all i would almost double the variances and covariances compared with the model with fixed $\mathbf{P}$. The correlations are virtually unchanged.

The Case $T = \infty$

The limiting variance–covariance vector, which we write as $\boldsymbol{\mu}_2$, satisfies

$$\boldsymbol{\mu}_2 = \boldsymbol{\mu}_2(E\mathbf{P} \times E\mathbf{P}) + \{\mathbf{n} - \mathbf{n}\mathbf{D}^{-1} + (\mathbf{n} \times \mathbf{n})\,\mathbf{J}'\mathbf{D}^{-1}\}\,\mathbf{R}_E, \tag{16}$$

where $\mathbf{n}$ is the vector of limiting expected grade sizes satisfying $\mathbf{n} = \mathbf{n}\mathbf{P}$ (assuming that $\mathbf{P}$ is irreducible). The effect of variation in $\mathbf{P}$ on $\boldsymbol{\mu}_2$ can, again, be most easily seen by considering values of D_i in relation to n_i. If, for example, we were to take $D_i = (n_i - 1)/K$ then (16) would become

$$\boldsymbol{\mu}_2 = \boldsymbol{\mu}_2(E\mathbf{P} \times E\mathbf{P}) + (1 + K)\,\mathbf{n}\mathbf{R}_E. \tag{17}$$

For fixed $\mathbf{P}$ the solution is given by setting $K = 0$. For $K > 0$ the solution will be exactly the same as in the fixed $\mathbf{P}$ problem with the total size of the system multiplied by $1 + K$. Now the solution of (17) with $K = 0$ gives

$$\operatorname{var}(n_j(\infty)) = n_j\{1 - (n_j/N)\}$$

and

$$\operatorname{cov}(n_j(\infty), n_l(\infty)) = -n_j n_l/N \quad (j, l = 1, 2, \ldots, k). \tag{18}$$

Thus if the D_i's were of the same order as the corresponding n_i's the limiting variances and covariances would be approximately doubled.

Numerical calculations

The foregoing arguments give a good general idea of the effects of a random $\mathbf{P}$ on errors of prediction. They show, in particular, that the importance of error from this source relative to the inherent statistical error depends on the size of the D_i's compared with the stock sizes. However, they do not tell us very much about the rate of approach to the limiting case or about what happens when the D_i's do not bear a constant ratio to the initial or final stocks. These points can be easily investigated by numerical calculation and an example is given in Table 1 relating to a system with seven states. The expected value of $\mathbf{P}$ used for these calculations is the Glass and Hall social mobility matrix given, for example, in Bartholomew (1973, Table 2.1); the states represent social classes, 1 being the highest and 7 the lowest:

	1	2	3	4	5	6	7
1	0·388	0·146	0·202	0·062	0·140	0·047	0·015
2	0·107	0·267	0·227	0·120	0·206	0·053	0·020
3	0·035	0·101	0·188	0·191	0·357	0·067	0·061
4	0·021	0·309	0·112	0·212	0·430	0·124	0·062
5	0·009	0·024	0·075	0·123	0·473	0·171	0·125
6	0·000	0·013	0·041	0·088	0·391	0·312	0·155
7	0·000	0·008	0·036	0·083	0·364	0·235	0·274

The initial class structure assumed for the calculation was $\mathbf{n}(0) = (100, 200, 300, 400, 500, 600, 700)$. For the above matrix the limiting class structure for a system of size 2,800 is $\mathbf{n} = (64{\cdot}4, 117{\cdot}6, 264{\cdot}4, 355{\cdot}6, 1{,}145{\cdot}2, 509{\cdot}6, 361{\cdot}2)$. Table 1 shows the variances for various assumptions about the D_i's. In the first three rows of each block

TABLE 1

Expected values and variances of stock numbers in a closed system governed by the Glass and Hall transition matrix with Dirichlet prior distributions having parameters $\mathbf{D} = (D_1, D_2, \ldots, D_k)$.

(For further details see text)

			Category						
T			1	2	3	4	5	6	7
1		$E\mathbf{n}(1)$	83·6	139·3	254·1	344·7	1,060·2	522·2	395·9
	$D_i =$	∞	65·6	118·8	219·4	295·5	641·2	402·2	318·4
		400	102·4	212·5	435·8	641·2	1,462·4	966·9	786·0
		50	359·5	868·6	1,950·7	3,061·5	7,210·4	4,919·5	4,059·2
	$\mathbf{D} =$	$\mathbf{n}$	157·1	276·7	489·8	650·8	1,438·8	921·9	758·5
		$\frac{1}{2}\mathbf{n}$	248·6	434·7	760·3	1,006·1	2,236·4	1,441·5	1,198·6
2		$E\mathbf{n}(2)$	73·0	123·9	250·1	352·7	1,129·1	508·3	362·9
	$D_i =$	∞	67·7	115·9	225·2	307·9	670·5	412·0	313·0
		400	118·8	237·5	532·3	781·3	1,756·1	1,089·7	829·5
		50	476·5	1,088·6	2,682·0	4,095·6	9,355·3	5,833·6	4,445·2
	$\mathbf{D} =$	$\mathbf{n}$	145·4	238·8	453·4	612·6	1,338·2	825·9	629·3
		$\frac{1}{2}\mathbf{n}$	223·1	361·7	681·6	917·3	2,006·0	1,239·8	945·6
5		$E\mathbf{n}(5)$	64·4	116·7	245·9	356·3	1,146·2	509·2	361·2
	$D_i =$	∞	62·9	111·8	224·3	311·0	677·0	416·5	314·6
		400	119·9	244·7	562·7	829·9	1,862·8	1,152·4	871·3
		50	519·0	1,174·9	2,931·9	4,462·1	10,163·6	6,303·2	4,768·6
	$\mathbf{D} =$	$\mathbf{n}$	125·2	221·8	445·4	618·6	1,348·4	830·8	627·3
		$\frac{1}{2}\mathbf{n}$	187·5	331·7	666·6	926·2	2,019·9	1,245·2	940·0
10		$E\mathbf{n}(10)$	63·4	115·9	245·1	356·2	1,147·3	510·1	361·9
	$D_i =$	∞	62·0	111·1	223·7	310·9	677·2	417·2	315·2
		400	118·9	244·1	562·9	831·4	1,867·3	1,156·4	874·5
		50	517·3	1,174·8	2,937·4	4,474·5	10,198·4	6,330·9	4,789·0
	$\mathbf{D} =$	$\mathbf{n}$	122·4	219·6	443·6	618·5	1,349·1	832·7	629·0
		$\frac{1}{2}\mathbf{n}$	182·8	328·2	663·5	926·0	2,021·0	1,248·3	942·9

the D_i's are assumed to be equal having the common value shown. In the last two rows the vector of D_i's is proportional to the vector of initial stocks.

The conclusions which emerge from the table may be summarized as follows:

(a) The approach to the limit is rapid though not necessarily monotonic.

(b) The qualitative effects are similar whether we make the D_i's equal or proportional to initial stocks.

(c) The D_i's have to be very large indeed before the randomness of **P** can be ignored.

This analysis clearly shows that uncertainty about the transition matrix can easily introduce errors into the predictions of a magnitude comparable with the statistical error arising from the stochastic assumptions of the model.

7. ESTIMATION ERROR: BAYESIAN TREATMENT

Estimation error arises not because the matrix **P** is changing but because its elements are estimated. However, if we approach the problem from a Bayesian viewpoint the estimation uncertainty will be expressed as a posterior distribution over the elements of **P**. The position is then identical mathematically to the specification problem already discussed with one important difference. We have supposed, hitherto, that **P** would have a different realized value in each time period whereas now its value is fixed but we are uncertain about what it is. The distinction does not arise when forecasting one step ahead since then only one realization is involved. When forecasting T steps ahead the equation corresponding to (3) will be

$$\boldsymbol{\mu}^*(T) = \boldsymbol{\mu}^*(0)\, E(\boldsymbol{\Pi}^{\mathrm{T}}) \tag{19}$$

which coincides with (3) only when $T=1$. The evaluation of $E\boldsymbol{\Pi}^{\mathrm{T}}$ for $T>1$ has not been investigated and we therefore confine the discussion here to one step ahead predictions.

Suppose that historical data on the flows is available with N_{ij} denoting the number of movements between i and j. Then the likelihood function is

$$\prod_{i=1}^{k}\prod_{j=1}^{k} p_{ij}^{N_{ij}}.$$

The conjugate prior density is the Dirichlet distribution. If we denote its parameters for the ith row by $\beta_{i1},\beta_{i2},\ldots,\beta_{ik}$, to avoid confusion with the α's used earlier, the posterior density for the ith row of **P** is

$$\pi(p_{i1},p_{i2},\ldots,p_{ik}) = \left[\Gamma\left\{\sum_{j=1}^{k}(\beta_{ij}+N_{ij})\right\}\Big/\prod_{j=1}^{k}\Gamma(\beta_{ij}+N_{ij})\right]\prod_{j=1}^{k} p_{ij}^{\beta_{ij}+N_{ij}-1}. \tag{20}$$

If we adopt the common Bayesian practice of putting $\beta_{ij}=0$ for all i and j to represent prior ignorance then (20) becomes identical with (8) when $\alpha_{ij}=N_{ij}$. We then have

$$D_i = \sum_{j=1}^{k} N_{ij}+1$$

so that D_i-1 is the total stock at risk in state i. If the p_{ij}'s were estimated from one period's data then these stocks would be of the same order as the $n_i(0)$'s and (15) shows that this would roughly double the variances and covariances. Conversely, estimation error will only be negligible if the number of periods over which historical data are available is large, say 10 or more. Similar results to those given in this section have been obtained by Keenay (1974).

8. Estimation Error: Frequentist Treatment

The frequentist approach to prediction error raises questions about the appropriate choice of reference set. Since the prediction is being made at time zero it is natural to make probability calculations conditional on $\mathbf{n}(0)$ and all other information available at that time. However, if $\mathbf{P}$ is estimated from stocks and flows from the same process prior to time zero, a thoroughgoing frequentist treatment would have to condition on these initial stocks treating $\mathbf{n}(0)$ as a random variable since its value is determined by past flows. This does not appear to be sensible when $\mathbf{n}(0)$ is already known and so we propose the following approach in which a different reference set is used for the statistical and estimation components of error.

If we have past flow data then, using the notation of the previous section, the maximum likelihood estimator of p_{ij} is

$$\hat{p}_{ij} = N_{ij}/N_i,$$

where $N_i = \sum_{j=1}^{k} N_{ij}$. If the matrix of these estimates is denoted by $\hat{\mathbf{P}}$, the predicted stock numbers at time 1 will be

$$\hat{\mathbf{n}}(1) = \mathbf{n}(0)\hat{\mathbf{P}}. \tag{21}$$

At time 0, when the prediction is made, $\hat{\mathbf{P}}$ and $\mathbf{n}(0)$ are known and therefore $\hat{\mathbf{n}}(1)$ is fixed. We are interested in the size of the vector difference $\mathbf{n}(1)-\hat{\mathbf{n}}(1)$ and this may be conveniently investigated by considering the matrix

$$\begin{aligned}\mathbf{M} &= E\{\mathbf{n}(1)-\hat{\mathbf{n}}(1)\}'\{\mathbf{n}(1)-\hat{\mathbf{n}}(1)\}\\ &= E\{\mathbf{n}(1)-E\mathbf{n}(1)\}'\{\mathbf{n}(1)-E\mathbf{n}(1)\}+\{E\mathbf{n}(1)-\hat{\mathbf{n}}(1)\}'\{E\mathbf{n}(1)-\hat{\mathbf{n}}(1)\}\\ &= \mathbf{M}_F+\mathbf{M}_E \text{ say.}\end{aligned}$$

The first term, $\mathbf{M}_F$, is simply the variance–covariance matrix of the category sizes when $\mathbf{P}$ is known and this is the statistical component of the prediction error. The second term is the additional contribution arising from estimation error. Its value is fixed but unknown and we require some means of being able to gauge its likely magnitude. Since

$$\hat{\mathbf{n}}(1)-E\mathbf{n}(1) = \mathbf{n}(0)\hat{\mathbf{P}}-\mathbf{n}(0)\mathbf{P} = \mathbf{n}(0)(\hat{\mathbf{P}}-\mathbf{P})$$

it follows that

$$\mathbf{M}_E = (\hat{\mathbf{P}}-\mathbf{P})'\mathbf{n}'(0)\mathbf{n}(0)(\hat{\mathbf{P}}-\mathbf{P}). \tag{22}$$

The Bayesian method is, in essence, to treat $\mathbf{P}$ as a random variable and so find the expectation of $\mathbf{M}_E$. From a frequentist point of view, the natural approach is to find the expectation by conditioning on the stocks from which the flows used in estimating $\hat{\mathbf{P}}$ have arisen. This is straightforward if the estimation is based on data which are independent of the process whose future is being predicted. In this case we can continue to treat $\mathbf{n}(0)$ as fixed and the determination of the expectation of $\mathbf{M}_E$ is then a simple matter. This situation is rather unlikely in practice so we shall return later to a discussion of the more difficult case whose present stocks are not independent of the estimated transition probabilities.

If **P** is estimated from data independent of the present process the expectation of $\mathbf{M}_E$ may be obtained as follows. The expectation of the (j, l)th element of $\mathbf{M}_E$ is easily found to be

$$\sum_{i=1}^{k} \sum_{i'=1}^{k} n_i(0)\, n_{i'}(0) \operatorname{cov}(\hat{p}_{ij}, \hat{p}_{i'l}).$$

But $\operatorname{cov}(\hat{p}_{ij}, \hat{p}_{i'l}) = 0$ when $i \neq i'$ because observed flows from different categories are independent. Hence the (j, l)th element of $\mathbf{M}_E$ has expectation

$$\sum_{i=1}^{k} n_i^2(0) \operatorname{cov}(\hat{p}_{ij}, \hat{p}_{il}). \tag{23}$$

The standard theory of multinomial sampling gives that

$$\left.\begin{aligned} \operatorname{var}(\hat{p}_{ij}) &= p_{ij}(1-p_{ij})/N_i \\ \operatorname{cov}(\hat{p}_{ij}, \hat{p}_{il}) &= -p_{ij}\, p_{il}/N_i \end{aligned}\right\} \quad (i = 1, 2, \ldots, k).$$

Substitution in (23) now provides the estimation component of the total prediction error. The statistical component of the error is easily found by letting $D_i \to \infty$ for all i in(15). This gives

$$\left.\begin{aligned} \operatorname{var}(n_j(1)) &= \sum_{i=1}^{k} n_i(0)\, p_{ij}(1-p_{ij}) \\ \text{and} \qquad & \\ \operatorname{cov}(n_j(1), n_l(1)) &= -\sum_{i=1}^{k} n_i(0)\, p_{ij}\, p_{il} \end{aligned}\right\} \quad (j, l = 1, 2, \ldots, k). \tag{24}$$

We may now make a comparison between the frequentist and Bayesian approaches to the error of one-step prediction with estimated transition probabilities. For simplicity we look at the variances, but similar results can be written down immediately for the covariances.

The frequentist expresses his uncertainty about the size of the jth state in terms of $E(n_j(1)-\hat{n}_j(1))^2$ which we have shown to be

$$\sum_{i=1}^{k} n_i(0)\, p_{ij}(1-p_{ij}) + \sum_{i=1}^{k} n_i^2(0)\, p_{ij}(1-p_{ij})/N_i.$$

However, he does not know the values of the p_{ij}'s so he must replace them by estimates. His estimate of the mean square prediction error will therefore be

$$\sum_{i=1}^{k} n_i(0) \frac{N_{ij}}{N_i}\left(1-\frac{N_{ij}}{N_i}\right) + \sum_{i=1}^{k} n_i^2(0) \frac{N_{ij}}{N_i^2}\left(1-\frac{N_{ij}}{N_i}\right) (j = 1, 2, \ldots, k). \tag{25}$$

The corresponding expression of uncertainty for the Bayesian will be $\operatorname{var}(n_j(1))$ as given in (15). This involves Ep_{ij} for $i = 1, 2, \ldots, k$. The posterior density of the p_{ij}'s with $\beta_{ij} = 0$ has $Ep_{ij} = N_{ij}/N_i$ so the variance is given by

$$\operatorname{var}(n_j(1)) = \sum_{i=1}^{k} n_i(0) \frac{N_{ij}}{N_i}\left(1-\frac{N_{ij}}{N_i}\right) + \sum_{i=1}^{k} \frac{n_i^2(0)+n_i(0)}{N_i+1} \frac{N_{ij}}{N_i}\left(1-\frac{N_{ij}}{N_i}\right) \quad (j = 1, 2, \ldots, k). \tag{26}$$

The comparison of (25) with (26) shows that they differ only in the second sum where the relative difference is of order $n_i(0)^{-1}$. For most practical purposes, therefore, the two approaches lead to the same measure of uncertainty and the choice between the two approaches can be regarded as a matter of taste.

As we have already noted it is not very likely that **P** will be estimated independently. It will usually be based on the history of the process itself which means that $\mathbf{n}(0)$ is at least partly determined by the flows which have been used to estimate **P**. In the extreme case when the transition probabilities have been estimated solely from flows over the interval from -1 to 0 then

$$\sum_{i=1}^{k} N_{ij} = n_j(0) \quad \text{or} \quad \sum_{i=1}^{k} N_i \hat{p}_{ij} = n_j(0) \quad (j = 1, 2, \ldots, k).$$

This raises the question of whether the expectation of $\mathbf{M}_E$ should be calculated by conditioning on the N_i's alone or on the N_i's and the $n_i(0)$'s. This is a matter which requires further investigation. If **P** is estimated from extensive stock and flow data extending over several time periods the dependence between $\hat{\mathbf{P}}$ and $\mathbf{n}(0)$ is not likely to be great and so the formulae given here will give a good idea of the size of the estimation error component. In any event they have the alternative justification provided by the Bayesian approach.

Since there is so little to choose between the Bayesian and frequentist methods for one-step prediction it would seem reasonable to adopt whichever method offers the easier way forward for predicting more than one step ahead. We have already commented on the practical difficulties with the Bayesian approach and matters are no easier using frequentist methods. Here we have to find $E(\hat{\mathbf{P}}^T)$ and evaluate expressions like (22) with $\hat{\mathbf{P}}^T$ in place of $\hat{\mathbf{P}}$. Although straightforward in principle the algebra is cumbersome and no simple way of handling it is obvious.

9. Extension of the Theory to Open Systems

There are two commonly used versions of the Markov model for open systems both described, for example, in Bartholomew (1973). In either case the internal transitions are governed by a transition matrix **P** but this no longer has row sums equal to one. In addition there is a loss probability associated with each category governing loss from the system. If we denote this by $p_{i,k+1}$ for the ith category then $\sum_{j=1}^{k+1} p_{ij} = 1$ $(i = 1, 2, \ldots, k)$. The two models differ in the assumptions made about input to the system. In the first, the inflow is specified as an aggregate number together with a vector of probabilities governing the allocation to categories. The total inflow may be fixed or it may be a random variable. We shall refer to this as the *given input* version of the open Markov model.

The second version, called the *given size* model, requires that the total size be kept fixed. In this case the total inflow is then determined by the total outflow but the allocation of new entrants still requires the specification of a "recruitment" vector.

Taking the given input model first it is clear that there is nothing in the derivation of (3), which required **P** to be a stochastic matrix. We can therefore take this as it stands and simply add on a term to account for the independent variability which arises from the input. This is the vector $\boldsymbol{\mu}_0(T)$ which appears in equation (3.32) of Bartholomew (1973).

The equation satisfied by the limiting second-order moments will now have the form

$$\boldsymbol{\mu}_2 = \boldsymbol{\mu}_2(E\mathbf{P} \times E\mathbf{P}) + \mathbf{x}\mathbf{R}_E + \boldsymbol{\mu}_{0,2},$$

where $\boldsymbol{\mu}_{0,2}$ is that part of $\boldsymbol{\mu}_0(\infty)$ containing the second-order moments and $\mathbf{x}$ is the same as the vector multiplier of $\mathbf{R}_E$ in (16).

The case of given size may be dealt with as follows. Suppose first that the total size is constant; then each loss will give rise to one gain. Thus the probability that a loss from i is associated with a gain to j is

$$q_{ij} = p_{ij} + r_j\, p_{i,k+1}, \tag{27}$$

where r_j is the probability that an entrant goes into j. The system may therefore be treated as if it were closed with stochastic transition matrix $\mathbf{Q} = \{q_{ij}\}$. This argument does not hold if new entrants are allocated in fixed proportions since then r_j in (27) is not a probability. The application of the theory for closed systems is now straightforward. All that is required is to convert the distribution of the p_{ij}'s into one for the q_{ij}'s. It follows at once from (27) that

$$\operatorname{cov}(q_{ij}, q_{il}) = \operatorname{cov}(p_{ij}, p_{il}) + r_j r_l \operatorname{var}(p_{i,k+1}) + r_j \operatorname{cov}(p_{il}, p_{i,k+1}) + r_l \operatorname{cov}(p_{ij}, p_{i,k+1}). \tag{28}$$

In the Dirichlet case, when $\operatorname{cov}(p_{ij}, p_{il}) = (\delta_{jl} Ep_{ij} - Ep_{ij} Ep_{il})/D_i$, this becomes

$$\operatorname{cov}(q_{ij}, q_{il}) = (\delta_{jl} Eq_{ij} - Eq_{ij} Eq_{il})/D_i - Ep_{i,k+1}\, r_j(\delta_{jl} - r_l)/D_i.$$

The last term on the right-hand side will be zero in the important practical case when $r_1 = 1, r_2 = r_3 = \ldots = r_k = 0$. In this case the matrix of the covariances of the q_{ij}'s has the same form as that of the p_{ij}'s.

Some illustrative calculations for the case of fixed input are given in Table 2.

TABLE 2

Expected values and variances of stock numbers in an open system with parameters as given in the text, $\mathbf{n}(0) = (126, 82, 27, 11)$ *and all recruitment into the bottom grade*

		Category			
T		1	2	3	4
	$E\mathbf{n}(1)$	127·6	83·2	27·1	10·6
1	$D_i = \{\ \infty$	47·1	23·6	6·5	1·8
	50	111·4	72·9	13·7	2·4
	$E\mathbf{n}(2)$	128·7	84·4	27·2	10·2
2	$D_i = \{\ \infty$	71·5	39·3	11·3	3·1
	50	170·1	120·8	23·8	4·2
	$E\mathbf{n}(3)$	129·5	85·5	27·3	9·9
3	$D_i = \{\ \infty$	84·2	50·5	14·9	4·2
	50	201·6	154·3	31·5	5·6
	$E\mathbf{n}(5)$	130·5	87·5	27·7	9·4
5	$D_i = \{\ \infty$	94·3	65·3	19·9	5·6
	50	228·2	197·4	42·4	7·5
	$E\mathbf{n}(\infty) = n$	131·6	93·5	31·4	8·3
∞	$D_i = \{\ \infty$	98·6	87·7	31·0	8·3
	50	241·9	264·3	70·6	12·1

The data on which the example is based are taken from Forbes (1971) and relate to the officer grade structure of one of the women's services. The transition matrix is

$$\mathbf{P} = \begin{bmatrix} 0{\cdot}715 & 0{\cdot}113 & 0 & 0 \\ 0 & 0{\cdot}841 & 0{\cdot}044 & 0 \\ 0 & 0 & 0{\cdot}869 & 0{\cdot}028 \\ 0 & 0 & 0 & 0{\cdot}894 \end{bmatrix}$$

and the loss probabilities are the complements of the row sums. It has been assumed that the D_i's are the same for each row and the table compares the cases $D = 50$ and $D = \infty$. The reason for choosing $D = 50$ is that the average grade size also has this value which should roughly double the variance. The recruitment is random with mean 37·5 and variance 21·4 and all entrants go into the bottom grade.

The conclusions to be drawn are very similar to those of the previous example. Uncertainty about **P** has a considerable effect and, when expressed in terms of the variances of the future grade sizes, is likely to be of the same magnitude as the statistical error.

ACKNOWLEDGEMENTS

The work on which this paper is based was begun while I was visiting Bell Laboratories at Holmdel, New Jersey. I derived considerable benefit from discussions with colleagues in the Operations Analysis Centre and, in particular, with Roshan Chadda and Luis Boza. I am also grateful to Gordon Keenay of the Civil Service Department for showing me his Ph.D. thesis containing the result on Bayesian estimation error. The problems associated with the choice of reference set in the frequentist approach to estimation error were drawn to my attention by Mr D. F. Jarrett who is currently investigating some of the questions raised in Section 7. I am grateful to the referees for the reference to Hoem's work and for suggestions which have improved the presentation.

REFERENCES

BARTHOLOMEW, D. J. (1973). *Stochastic Models for Social Processes*, 2nd ed. Chichester: Wiley. (See also 1st ed., 1967.)

COLEMAN, J. S. (1964). *Introduction to Mathematical Sociology*. London: Collier–Macmillan.

FORBES, A. F. (1971). Markov chain models for manpower systems. In *Manpower and Management Science* (D. J. Bartholomew and A. R. Smith, eds), pp. 93–113. London: English Universities Press; Lexington, Mass.: D. C. Heath and Co.

GRAYBILL, F. A. (1969). *Introduction to Matrices with Applications in Statistics*. Belmont, California: Wadsworth.

HOEM, J. M. (1973). *Levels of Error in Population Forecasts*, Artikler fra Statistisk Sentralbyrå, Nr. 61, Oslo.

KEENAY, G. A. (1974). Manpower planning in large organizations: a statistical approach. Ph.D. Thesis, University of Cambridge.

MARTIN, J. J. (1967), *Bayesian Decision Problems and Markov Chains*. New York: Wiley.

POLLARD, J. H. (1966). On the use of the direct matrix product in analysing certain stochastic population models. *Biometrika*, **53**, 397–415.

—— (1973). *The Mathematical Theory of the Growth of Populations*. Cambridge: University Press.

SALES, P. (1971). The validity of the Markov chain model for a branch of the Civil Service. *Statistician*, **20**, 85–110.

STAFF, P. J. and VAGHOLKAR, M. K. (1971). Stationary distributions of open Markov processes in discrete time with application to hospital planning. *J. Appl. Prob.*, **8**, 668–680.