ANDREI ROGERS AND ROBERT MILLER

Estimating a Matrix Population Growth Operator From Distributional Time Series

Center for Planning & Development Research: Institute of Urban & Regional Development
University of California, Berkeley

Reprint No. 34

The Center for Planning and Development Research is a unit within the Institute of Urban and Regional Development. It is devoted to the study of a wide range of activities in environmental design and urban and regional development. Researchers in the Center are engaged in investigating the processes of urban and regional growth, the consequences of alternative patterns of growth, and the scientific bases for development policy and planning.

Toward these ends, current research is directed to improving theories and methods of planning, to building models capable of simulating urban growth processes, and to formulating improved criteria for urban and regional development policies. Fields of interest to the Center include the development of new and established communities, housing and renewal programming, the effects of technological development on cities and urban life, metropolitan planning, the economics of city-building in the United States and in developing nations, and the social planning and human development issues that accompany urbanization.

The Center contains the Architectural Design Laboratory, which is devoted to research in building technology, the behavioral bases for design, and design processes.

Estimating a Matrix Population Growth Operator From Distributional Time Series

ANDREI ROGERS AND ROBERT MILLER
University of California, Berkeley

Reprinted from ANNALS OF THE ASSOCIATION OF AMERICAN GEOGRAPHERS
Vol. 57, No. 4, December, 1967, pp. 751–756
Made in United States of America

ESTIMATING A MATRIX POPULATION GROWTH OPERATOR FROM DISTRIBUTIONAL TIME SERIES

ANDREI ROGERS AND ROBERT MILLER
University of California, Berkeley

ABSTRACT. Most efforts to analyze and forecast interregional population growth and distribution are hampered by the general paucity of reliable data on interregional migration. This has led demographers to rely on crudely constructed measures of net migration. Recent efforts to express the population growth process in matrix form, however, suggest a means whereby the schedule of growth that operates in an interregional system may be estimated solely on the basis of historical data on interregional population distributions. Such a method is described and illustrated in this paper. The method estimates a growth regime which, when applied to the initial interregional population distribution in the time series, most closely reproduces the remaining population distributions in the historical sequence.

A common constraint in interregional population analysis and forecasting is the absence of reliable data concerning the behavior of the fundamental components of population change, particularly in- and out-migration. Thus considerable importance attaches to possible ways of circumventing this problem. Toward this end, the recent literature on matrix models of interregional population growth is especially relevant.[1] Together with the now well developed techniques for estimating the transition matrices of Markov chains,[2] these models suggest a method whereby the growth regime of an interregional population system may be estimated solely on the basis of a history of interregional population distributions. In this paper we present such a method and illustrate it using data on a two-region system consisting of California and the rest of the United States.

THE COMPONENTS-OF-CHANGE POPULATION MODEL

Current methods of population analysis and forecasting exhibit a large number of variations. By and large, however, one may distinguish between those forms which involve only projections of total births, deaths, and migration, from those which focus on age and sex disaggregated cohorts. The latter class commonly are referred to as cohort-survival models, the former may be defined as components-of-change models.

The component method of population forecasting begins by projecting each of the major components of population change, i.e., births, deaths, and net migration. These components

Accepted for publication September 28, 1966.

[1] N. Keyfitz, "Matrix Multiplication as a Technique of Population Analysis," *Milbank Memorial Fund Quarterly* (1964), pp. 68–83, and "The Population Projection as a Matrix Operator," *Demography*, Vol. 1 (1964), pp. 56–73; A. Rogers, "Matrix Methods of Population Analysis," *Journal of the American Institute of Planners*, Vol. 32 (1966), pp. 40–44; "A Note on the Temporal Decomposition of Interpoint Transition Matrices," *Journal of Regional Science*, Vol. 6 (1966), pp. 53–56; "The Multiregional Matrix Growth Operator and the Stable Interregional Age Structure, *Demography* (Fall 1966), forthcoming; "Matrix Analysis of Interregional Population Growth and Distribution," *Papers and Proceedings of the Regional Science Association*, VIth European Congress, Vienna (August 1966), pp. 537–44.

[2] T. C. Lee, G. G. Judge, and T. Takayama, "On Estimating the Transition Probabilities of a Markov Process," *Journal of Farm Economics*, Vol. 43 (1965), pp. 742–62; A. Madanksy, "Least Squares Estimation in Finite Markov Processes," *Psychometrika*, Vol. 25 (1959), pp. 137–44; G. A. Miller, "Finite Markov Processes in Psychology," *Psychometrika*, Vol. 17 (1952), pp. 149–67; L. G. Telser, "Least Squares Estimates of Transition Probabilities," in F. Christ, (Ed.) *Measurement in Economics* (Palo Alto, California: Stanford University Press, 1963), pp. 272–92; H. Theil and G. Rey, "A Quadratic Programming Approach to the Estimation of Transition Probabilities," *Management Science*, Vol. 12 (1966), pp. 714–21.

TABLE 1.—COMPONENTS OF POPULATION CHANGE IN CALIFORNIA AND THE REST OF THE UNITED STATES 1955–1956

(1) Region	(2) 1955 population (000's)	(3) Births 1955–56 (000's)	(4) Deaths 1955–56 (000's)	(5) Net migration 1955–56 (000's)	(6) Crude birth rate	(7) Crude death rate	(8) Crude net migration rate	(9) 1956 population (000's)
California	12988	315	103	208	.0242	.0079	.0160	13408
Rest of the United States	152082	3761	1434	–208	.0247	.0094	–.0014	154201

Source: U.S. Bureau of the Census, *Statistical Abstract of the United States: 1962* (Eighty-third edition) (Washington, D.C.: Government Printing Office, 1962), and California Department of Finance, Financial and Population Research Section, *California Migration 1955–1960* (Sacramento, California: 1964). Data adjusted by authors.

are then combined with current population data to provide a future population estimate. Symbolically, the relationship between these variables may be expressed by the equation

$$(1) \qquad w^{(t+1)} = w^{(t)} + b^{(t)} - d^{(t)} + n^{(t)}$$

where $w^{(t)}$ = population at time t;
$b^{(t)}$ = number of births between t and $t+1$;
$d^{(t)}$ = number of deaths between t and $t+1$;
$n^{(t)}$ = number of net migrants between t and $t+1$.

Typically, b, d, and n are derived for each interval of time by the application of crude birth, death, and net migration rates so that

$$(2) \qquad \begin{aligned} w^{(t+1)} &= w^{(t)} + \beta w^{(t)} - \delta w^{(t)} + \eta w^{(t)} \\ &= (1+\beta-\delta+\eta)w^{(t)} = g w^{(t)} \end{aligned}$$

where β = crude birth rate;
δ = crude death rate;
η = crude net migration rate;
g = the "growth multiplier."

Consider, for example, the data in Table 1. Given a base population and the birth, death, and net migration rates for a region such as California, one can by means of Equation (2) estimate the population of the region a unit time interval hence. Thus, for California, we have (in 000's):

$$\begin{aligned} w^{(1956)} &= (1 + .0242 - .0079 + .0160)w^{(1955)} \\ &= (1.0323)\,12{,}988 \\ &= 13{,}408\,. \end{aligned}$$

In an analogous manner, we may derive the 1956 population for the rest of the United States which is presented in column 9 of Table 1.

THE MATRIX INTERREGIONAL COMPONENTS-OF-CHANGE POPULATION MODEL

Consider now the matrix representation of Equation (2) for an interregional system:

$$(3) \qquad w^{(t+1)} = (I + B - D + N)w^{(t)} = Gw^{(t)}$$

where $w^{(t)}$ = a column vector whose i^{th} element denotes the population of region i at time t;
I = an identity matrix;
B, D, N = diagonal matrices whose nonzero elements denote the crude birth, death, and net migration rates of the m regions;
$G = (I + B - D + N)$ = the matrix "growth operator."

Equation (3) now may be used to rederive the 1956 populations found in Table 1 (in 000's):

$$w^{(1956)} = \begin{bmatrix} 1.0323 & .0 \\ .0 & 1.0139 \end{bmatrix} \begin{bmatrix} 12{,}988 \\ 152{,}082 \end{bmatrix} = \begin{bmatrix} 13{,}408 \\ 154{,}201 \end{bmatrix}.$$

With the matrix formulation of the interregional components-of-change population model, one can, without added effort, focus on in- and out-migration flows instead of net migration totals. One merely introduces a migration transition matrix, P, whose elements, p_{ij}, describe the proportion of people who, during a specified time period, moved from region i to region j. Such a transition matrix has been computed for our two-region system and appears in Table 2.

Given a transition matrix P, one may replace, with its transpose P', the matrices I and N in Equation (3) to define the interregional population model

$$(4) \qquad w^{(t+1)} = (B - D + P')w^{(t)} = Gw^{(t)}$$

where $G = (B - D + P')$ = the matrix "growth operator."

With Equation (4) we again can solve for the 1956 populations in Table 1:

TABLE 2.—MIGRATION TRANSITION MATRIX: 1955–56 (IN 000'S)

From	To: California	To: U.S.	1955 total
California	12,831	157	12,988
	(.9879)	(.0121)	(1.0000)
Rest of the United States	365	151,717	152,082
	(.0024)	(.9976)	(1.0000)

Source: A. Rogers, "A Note on the Temporal Decomposition of Interpoint Transition Matrices," *Journal of Regional Science*, Vol. 6, No. 2 (Fall 1966).

$$w^{(1956)} = \begin{bmatrix} 1.0042 & .0024 \\ .0121 & 1.0129 \end{bmatrix} \begin{bmatrix} 12,988 \\ 152,082 \end{bmatrix} = \begin{bmatrix} 13,408 \\ 154,201 \end{bmatrix}.$$

If we assume that the growth regime for this interregional system will remain constant over a 5-year interval, we may forecast the sequence of interregional distributions presented in Table 3 by repeatedly applying Equation (4).

The projected 1960 population figures are slightly lower than those reported by the U.S. Census. However, they are adequate for our expositional purposes and will be used to demonstrate the estimation techniques.

ESTIMATING THE MATRIX GROWTH OPERATOR

Consider now the growth process for our two-region example, over two sequential periods, and assume that the growth regime remains unchanged during this time interval. We have then:

$$\begin{bmatrix} w_1^{(t+1)} \\ w_2^{(t+1)} \end{bmatrix} = \begin{bmatrix} g_{11} & g_{21} \\ g_{12} & g_{22} \end{bmatrix} \begin{bmatrix} w_1^{(t)} \\ w_2^{(t)} \end{bmatrix}$$

and

$$\begin{bmatrix} w_1^{(t+2)} \\ w_2^{(t+2)} \end{bmatrix} = \begin{bmatrix} g_{11} & g_{21} \\ g_{12} & g_{22} \end{bmatrix} \begin{bmatrix} w_1^{(t+1)} \\ w_2^{(t+1)} \end{bmatrix}.$$

Assume that the w_i's have been observed and therefore are known, but that the matrix growth operator, G, is unknown and has to be estimated solely on the basis of the observed w_i's. We have, then, a problem in the solution of a system of 4 equations in 4 unknowns. Thus, for example, given $w^{(1955)}$, $w^{(1956)}$, and $w^{(1957)}$ of Table 3, we may solve the system described below and "retrieve" the matrix G:

$$\begin{aligned} 12,988\, g_{11} + 152,082\, g_{21} &= 13,408 \\ 12,988\, g_{12} + 152,082\, g_{22} &= 154,201 \\ 13,408\, g_{11} + 154,201\, g_{21} &= 13,834 \\ 13,408\, g_{12} + 154,201\, g_{22} &= 156,352 \end{aligned}$$

whence,

$$g_{11} = 1.0042 \qquad g_{21} = .0024$$
$$g_{12} = .0121 \qquad g_{22} = 1.0129$$

It should be noted that, by solving the equation system, we have extracted migration information from data only on interregional population distributions.

Consider next, the more general problem of estimating an operator G which, when applied to the vector $w^{(1955)}$, reproduces, as accurately as possible, the entire time series in Table 3. If, as before, we assume that G does not change during the 5-year interval, we have a system of 10 equations in 4 unknowns. Thus a criterion of goodness-of-fit needs to be introduced in order to arrive at a unique solution. Three have been suggested in the literature: 1) the unrestricted least-squares estimator;[3] 2) the minimum absolute deviations estimator;[4] and 3) the restricted least-squares estimator.[5] The first adopts the well-known estimation technique traditionally used in regression analysis; the latter two require a mathematical programming formulation. In this paper we will consider only the first two estimation techniques. The restricted least-squares solution will be described in a forthcoming paper.

[3] Madanksy, *op. cit.*, footnote 2; Miller, *op. cit.*, footnote 2.

[4] V. Ashar, and T. D. Wallace, "A Sampling Study of Minimum Absolute Deviations Estimator," *Operations Research*, Vol. 11 (1963), pp. 747–51; W. D. Fisher, "A Note on Curve Fitting with Minimum Deviations by Linear Programming," *Journal American Statistical Association*, Vol. 56 (1956), pp. 359–62; O. J. Karst, "Linear Curve Fitting Using Least Deviations," *Journal American Statistical Association*, Vol. 53 (1958), pp. 118–32; Lee, et al., *op. cit.*, footnote 2.

[5] G. G. Judge, and T. Takayama, "Inequality Restrictions in Regression Analysis," *Journal American Statistical Association*, Vol. 61 (1966), pp. 166–81; Lee, et al., *op. cit.*, footnote 2; Theil and Rey, *op. cit.*, footnote 2.

TABLE 3.—DISTRIBUTIONAL TIME SERIES GENERATED BY A CONSTANT GROWTH OPERATOR: CALIFORNIA AND THE REST OF THE UNITED STATES, 1955–1960

Region	1955	1956	1957	1958	1959	1960
California	12,988	13,408	13,834	14,267	14,707	15,155
Rest of the United States	152,082	154,201	156,352	158,536	160,754	163,006
Total	165,070	167,609	170,186	172,803	175,461	178,161

Source: Computed by authors.

The Unrestricted Least-Squares Estimator

We begin by assuming that the observed vectors, w, denoting the population in each of m regions, are generated by the linear statistical model

$$w^{(t+1)} = Gw^{(t)} + u^{(t)} \tag{5}$$

where $u^{(t)}$ is a $(m \times 1)$ vector of error terms. Expressing Equation (5) in the more familiar linear regression form, we have:

$$y = Xg + u \tag{6}$$

where, for our two-region example,

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \quad g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}$$

or, more specifically,

$$\underset{2n \times 1}{y} = \begin{bmatrix} w_1^{(t+1)} \\ w_1^{(t+2)} \\ \cdot \\ \cdot \\ w_1^{(t+n)} \\ \hline w_2^{(t+1)} \\ w_2^{(t+2)} \\ \cdot \\ \cdot \\ w_2^{(t+n)} \end{bmatrix} \quad \underset{2^2 \times 1}{g} = \begin{bmatrix} g_{11} \\ g_{21} \\ \hline g_{12} \\ g_{22} \end{bmatrix}, \quad \underset{2n \times 1}{u} = \begin{bmatrix} u_{11} \\ u_{12} \\ \cdot \\ \cdot \\ u_{1n} \\ \hline u_{21} \\ u_{22} \\ \cdot \\ \cdot \\ u_{2n} \end{bmatrix}$$

and

$$\underset{2n \times 2^2}{X} = \left[\begin{array}{cc|cc} w_1^{(t)} & w_2^{(t)} & 0 & 0 \\ w_1^{(t+1)} & w_2^{(t+1)} & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ w_1^{(t+n-1)} & w_2^{(t+n-1)} & 0 & 0 \\ \hline 0 & 0 & w_1^{(t)} & w_2^{(t)} \\ 0 & 0 & w_1^{(t+1)} & w_2^{(t+1)} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & w_1^{(t+n-1)} & w_2^{(t+n-1)} \end{array}\right]$$

Recalling that the least-squares estimator of g is

$$\hat{g} = (X'X)^{-1}X'y, \tag{7}$$

we may apply Equation (7) to our two-region example and once again retrieve the growth matrix G:

$$\hat{G} = \begin{bmatrix} \hat{g}_{11} & \hat{g}_{21} \\ \hat{g}_{12} & \hat{g}_{22} \end{bmatrix} = \begin{bmatrix} 1.0042 & .0024 \\ .0121 & 1.0129 \end{bmatrix}$$

Unrestricted least-squares estimation, of the kind described above, leaves unresolved a serious problem in the estimation of growth operators. This is the possible entry of negative elements into G. These, of course, are by definition inadmissible estimates. To cope with this problem, Telser[6] has suggested a method whereby negative elements are set equal to zero and the matrix is then adjusted to compensate for this change. Such a procedure is unsatisfactory for our purposes, however, since setting a g_{ij} equal to zero would imply either: 1) zero migration from i to j, or 2) the extinction of the population in region i. Thus, for a more reasonable approach, one must adopt a method which incorporates a nonnegativity constraint in the estimation process. Such a restriction may be introduced by a mathematical programming formulation of the estimation problem.

The Minimum Absolute Deviations Estimator

The minimum absolute deviations estimation technique uses linear programming to find that vector which minimizes the sum of the absolute deviations between observed and predicated values. Thus we seek the vector g which minimizes

$$|y - Xg|'l \tag{8}$$

subject to the constraints

$$y = Xg + u \tag{9}$$

$$g \geqq 0. \tag{10}$$

[6] Tesler, *op. cit.*, footnote 2.

TABLE 4.—LINEAR PROGRAMMING TABLEAU FOR THE MINIMUM ABSOLUTE DEVIATIONS ESTIMATE

Objective	g_1'	g_2'	$\ldots$	g_r'	a_1'	a_2'	$\ldots$	a_r'	b_1'	b_2'	$\ldots$	b_r'	B
–1	$0'$	$0'$	$\ldots$	$0'$	$1'$	$1'$	$\ldots$	$1'$	$1'$	$1'$	$\ldots$	$1'$	0
	X_1				I_n				$-I_n$				y_1
		X_2				I_n				$-I_n$			y_2
			$\ldots$				$\ldots$				$\ldots$		$\ldots$
				X_r				I_n				$-I_n$	y_r

and $X_1 = X_2 = \ldots = X_r$.

Source: Lee, Judge, and Takayama, cited in footnote 2.

Where l is a unit vector of order $rn \times 1$, y_j is a $n \times 1$ vector of observations denoting the number of persons in region j at successive points in time, and n is the number of transition periods. X_j is a $n \times r$ matrix of realized values of the number of persons in each region at times t through $t + n - 1$, r is the number of regions, g_j is a $r \times 1$ vector of unknown elements of the growth matrix, and u_j is a $r \times 1$ vector of error terms.

In order for the problem to be solvable with linear programming solution methods, nonnegative variables are introduced for u such that

$$u = a - b \tag{11}$$

where a and b are vectors of order rn. One now may redefine the problem to be the minimization of

$$(a + b)'l \tag{12}$$

subject to

$$y = Xg + u = Xg + [I, -I]\begin{bmatrix} a \\ b \end{bmatrix} \tag{13}$$

and the usual nonnegativity restraints. The problem can be solved by the simplex method using the tableau presented in Table 4. Since the vectors a and b are linearly dependent, components of one or the other, but not both, will enter the solution. Thus, there are $r^2 + rn$ unknowns, but only rn constraints.

Applying the simplex method to the data in Table 3, we find the solution presented in Table 5. The minimum absolute deviations estimate of the growth matrix is:[7]

$$\hat{G} = \begin{bmatrix} 1.0044 & .0023 \\ .0131 & 1.0128 \end{bmatrix}.$$

[7] Theoretically we should have retreived exactly the G that was used to generate the time series. The slight errors are due to rounding.

TABLE 5.—LINEAR PROGRAMMING TABLEAU AND SOLUTION FOR THE TWO-REGION EXAMPLE

Input:

obj	g_{11}	g_{21}	g_{12}	g_{22}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	b_{11}	b_{12}	b_{13}	b_{14}	b_{15}	b_{21}	b_{22}	b_{23}	b_{24}	b_{25}	B
–1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	12,988	152,082			1										–1										13,408
	13,408	154,201				1										–1									13,834
	13,834	156,352					1										–1								14,267
	14,267	158,536						1										–1							14,707
	14,707	160,754							1										–1						15,155
			12,988	152,082						1										–1					154,201
			13,408	154,201							1										–1				156,352
			13,834	156,352								1										–1			158,536
			14,267	158,536									1										–1		160,754
			14,707	160,754										1										–1	163,006

Output:

Variable	Value	Position in basis
objective	2.6709	1
a_{11}	0.8970	2
g_{21}	0.0023	3
g_{11}	1.0044	4
b_{14}	0.1092	5
a_{15}	0.7797	6
a_{21}	0.6595	7
g_{22}	1.0128	8
b_{23}	0.1472	9
a_{24}	0.1902	10
g_{12}	0.0131	11

$$\hat{G} = \begin{bmatrix} \hat{g}_{11} & \hat{g}_{21} \\ \hat{g}_{12} & \hat{g}_{22} \end{bmatrix} = \begin{bmatrix} 1.0044 & .0023 \\ .0131 & 1.0128 \end{bmatrix}$$

Source: Computed by authors.

CONCLUSION

This paper has demonstrated that it is possible to estimate the structure of interregional population growth and migration, solely on the basis of data on interregional population distributions over several time periods. Moreover, it appears that the linear programming formulation of the estimation procedure provides a satisfactory method for approximating an unknown interregional regime of growth. Several problems may arise, however, when observed, rather than artificial, data are used. The assumption of a constant growth operator, for example, becomes less viable as the number of time periods is increased. This assumption may lead to a poor fit and force a large number of error terms into the solution basis, thereby forcing out some of the elements of the operator G. In tests with data generated by a changing G it was found that often an entire row of the estimated growth operator would be missing. Another problem occurs when the regions differ considerably in size. In such instances, an error in estimating a g_{ij} associated with a small region should not, perhaps, have the same weight as an identical error associated with the estimation of a g_{ij} associated with a large region. Hence, a weighted estimation procedure may need to be employed. Finally, the minimum absolute deviations principle may not be the best criterion to use for deriving the estimator. A more commonly adopted criterion is that of least-squares. Thus a quadratic programming approach, based on the least-squares principle, may provide improved results.