THE DETERMINATION OF A BASE POPULATION FOR COMPUTING MIGRATION RATES*

RALPH THOMLINSON**

I. INTRODUCTION

ALTHOUGH demographers have been computing various kinds of rates for fertility and mortality for some years, it is only recently that there have been many serious attempts to calculate rates of migration.[1] This weakness in demographic investigation was identified in Dudley Kirk's presidential address delivered to the Population Association of America on May 6, 1960:

> The study of internal migration is the stepchild of demography. Too little attention has been given by the leadership of our profession to the theory and measurement of migration, despite its role as the chief determinant of differences in population change and structure among local populations, and indeed now for many states. In the words of one leading authority in this field, the majority of recent migration studies are 'planlessly empirical and trivial in content.' This is a harsh judgment. The study of migration presents peculiar problems in terms of definition and complexity, but I feel confident that the application of the technical virtuosity so evident in the field of natality could yield great progress. It is our responsibility as demographers not to ignore the crucial problems because the data do not lend themselves readily to pat generalizations or to neat mathematical models. It is in this area of migration, with its complex variables of origin, destination, distance and time that the new computers may make the greatest contribution to demographic analysis.[2].

* Professor Georges Sabagh of the University of Southern California supplied valuable criticism of a draft of this paper.

** Department of Sociology, Los Angeles State College.

[1] See for example Donald J. Bogue, Henry S. Shryock, Jr., and Siegfried A. Hoermann, SUBREGIONAL MIGRATION IN THE UNITED STATES, 1935–40, Vol. I: Streams of Migration Between Subregions, (Scripps Foundation Studies in Population Distribution No. 5, [Oxford, Ohio, 1957]), especially pp. 7–14 and 49–50.

[2] Dudley Kirk, Some Reflections on Amercian Demography in the Nineteen Sixties, POPULATION INDEX, Vol. XXVI (October, 1960), p. 307.

Crude migration rates are much more troublesome than crude birth and death rates, for two reasons. First, the numerator (number of migrants) is harder to obtain; a thorough explanation of this situation is long and is not directly relevant to this paper.[3] Second, the denominator (the base population) is more difficult to ascertain, largely because migration involves two geographic entities rather than one and because a sometimes lengthy time period is involved. Scholars have not reached agreement concerning the most appropriate population base to be used as a denominator in the fraction expressing a rate of migration. This article contains an exposition of the qualities demanded for such a base population, followed by an evaluation of the most likely methods for determining the base population.

In order to make the procedures applicable to as many countries as possible, this paper makes minimal assumptions concerning the character of migration data. Since only a few countries have continuous migration registers, it seems preferable not to assume such a continuing count. In other countries, migration data are obtained a) through enumerating recent residential changes, b) by special origin-destination studies, c) through estimates based on survival ratios, or d) by computing net migration as a residual after other growth components are removed from total increase. The reasoning in this paper applies to any situation for which one can obtain data of the nature described in the following paragraph.

A migrant is defined here as a person who resides in a place different at the end of a specified period of time from the beginning of the time period. The time span is called the migration interval. Thus, "reverse movers" who make cancelling "in" and "out" moves of opposite direction during the interval are not recorded as migrants. This includes two types of people: those moving out of the area and then returning before the period concludes and those moving into the area

[3] Seven major factors contributing to the difficulty of defining and measuring migration are discussed *in* Ralph Thomlinson, Methodological Needs in Migration Research, *Population Review* [Madras, India] Vol. VI (January, 1962), pp. 59–64. Minor involvements are mentioned later in this paper.

and then leaving before the end of the period. Unless a country maintains a usable continuous register of migrants, these people will not ordinarily even be known to have moved. Persons moving entirely within the given place (usually politically defined) are "movers" but not "migrants."

II. Criteria for Base Populations in Migration Rates

In the establishment of a base population for computing demographic rates, four criteria must be satisfied: 1) the base population must correspond exactly to the population exposed to the event in question; 2) the data must be customarily available; 3) the data must be accurate (or, as in certain cases mentioned below, the magnitude and direction of error must be known); and 4) computation of the base must not be too cumbersome. Traditional usage is sometimes a criterion for selection of a base population, possibly because it is rarely necessary to defend use of traditional procedures, whereas abandonment of tradition must be defended.

The first criterion is critical. A base population must include all people who are exposed to the event being measured, and no other people. This is surprisingly difficult to ascertain, and it becomes more of a problem as the time period lengthens —especially when the interval is longer than a year.

In computing rates of migration, the following five groups of people are of concern: a) people who remain in the area during the entire migration interval, b) those born in the area during the interval, c) those dying in the area during the interval, d) those moving into the area during the interval, e) those moving out of the area during the interval. The first three are easy to manage; the last two pose intricate statistical problems.

The base population should be the average number of people in the area during the given period but should not include any people born or dying during the period, because people who were not alive throughout the entire migration period could not be recorded as migrants over the given migration interval. To illustrate, let us assume that migration statistics were tabulated from respondents' replies to the question, asked on April 1, 1960: "In what place did you live on April 1, 1955?" This means of acquiring data results in the operational definition of a migrant as a person reported to have been living in a place on April 1, 1955 which was different from his residence on April 1, 1960.[4] This definition requires that the person be alive at both dates. People dying during the period could never be asked about moving because they were not alive at the time of the 1960 interview; people born during the period had no 1955 residence.

Migrants themselves create difficulties. Ideally, no person should be a part of more than one base population.[5] If a migrant were included in the base population of every area in which he resided during the migration period, he might be counted many times;[6] fortunately, this undesirable situation is easily avoided by excluding migrants from the base populations of all areas except those in which they resided at the beginning or end of the migration interval. This dictum still leaves the possibility of including a migrant in two base populations: that of the area of his residence in 1955 and that of the area of his 1960 residence. In order to avoid the distortion accompanying this eventuality, we must choose one of these two residence-dates. The beginning date (in our example, 1955) appears preferable, because migration tables are based on initial residence (e.g., 1955) as the foundation for determining exposure to the possibility of becoming a migrant.

[4] Hereafter, to simplify the discussion, "1955" and "1960" will be used in preference to the more precise dates.

[5] From the standpoint of any one given area, it is perfectly legitimate (because totally irrelevant) for some people to be included also in another base population. However, when two or more rates are compared, proper comparison is made difficult by multiple inclusions. Since the person computing rates cannot be sure that the rates will never be used for purposes of contrast with other rates (such use being very common), it seems appropriate to prepare for this eventuality.

[6] Inclusion of one migrant in the base populations of two or more countries would result in artificially low migration rates (because of inflation of the denominators). This distortion of the rates increases in direct relation to increases in the frequency of moves made by multiple migrants—i.e., migrants making more than one move apiece.

If we take this position, people moving into an area after 1955 should not be included in the base population. Residents of an area in 1955 who moved out of the area between 1955 and 1960 are included in the migration table (provided they remained out of the area); hence, they should be a part of the base population. Persons moving away from the area between 1955 and 1960 who did not reside in the area in 1955 would not be recorded in the migration statistics; the base population should not be related to these out-migrants in any way.

The above discussion turns on the logic of who is exposed to moving—the first criterion listed above. As for the second criterion, any area having origin-destination data will also have adequate base population data. Regarding accuracy of the data, figures taken directly from an official census are ordinarily superior to inter-censal estimates, particularly for small areas. Fourth, simple addition and subtraction are easier to perform than interpolation; the difference is meaningful (although not highly important) when the operations are to be undertaken for a large number of areas.

Surprisingly, it is not requisite that a base population be precisely accurate. A base population that has the identical percentage and direction of error as the corresponding migration figure may be fully adequate. Since a migration rate customarily is defined as the number of migrants divided by the size of the base population, what matters is the relative proportions of people involved. If both the base population and the migration total used in the computation are four per cent too large, then the ratio of the two will be correct. However, it is extremely improbable that one will locate a set of such data that are all in error by the same amount and the same direction.

A question may be raised as to the applicability of these considerations to terminal areas as well as to source areas. Although the reasoning concerning migration rates for areas of origin is clear, it is difficult to conceptualize a base population for an area of destination; in one sense, it is everyone not in the terminal area itself. However, reflecting upon the fact that every geographic area enters into the migration stream both as a sending and as a receiving area, one is encouraged to form a tentative conclusion that the base population for an area should remain constant, regardless of whether it is treated as a source or as a terminal area.

Depending on the focus of interpretation, migration rates may require in the denominator either one or two numbers.[7] For simplicity, the above analysis has referred explicitly to the single-number situation. When an investigator wishes to use the product of two base populations (those of the area of origin and the area of destination)[8] the same criteria apply.

To return to our 1955–1960 example, what do we know about the qualities our base population should have in order to satisfy criterion 1) above? First, it must not include persons born after 1955. Second, it must not include persons dying before 1960. Third, it must include no in-migrants. Fourth, it must include only those out-migrants who resided in the area in 1955.

III. Evaluation of Possible Base Populations

Seven figures need to be considered as possible base populations for a given area and a migration period of *n* years: (a) the final population, (b) the initial population, (c) the mid-point population, (d) the final population *n* years of age and over, (e) the arithmetic mean of i—the final population aged *n* years and over, and ii—the initial population minus deaths occurring during the period to initial residents of the area, (f)

[7] Two base populations are necessary when using a gravitational approach or when measuring the stream of movement—i.e., when emphasis is on the move rather than on an area. A single base may suffice when the purpose is to compare the relative attractiveness of different areas or to assess the impact of migration on an area.

[8] See for instance John Q. Stewart, The 'Gravitation' or Geographic Drawing Power of a College, *Bulletin of the American Association of University Professors,* Vol. XXVII (February, 1941), pp. 70–75; John Q. Stewart, A Measure of the Influence of a Population at a Distance, *Sociometry,* Vol. V (February, 1942), pp. 63–71; George K. Zipf, "The P_1P_2/D Hypothesis: On the Intercity Movement of Persons, *American Sociological Review,* Vol. XI (December, 1946), pp. 677–685; Fred Charles Iklé, Sociological Relationship of Traffic to Population and Distance, *Traffic Quarterly,* Vol. VIII (April, 1954), pp. 123–136; and Theodore R. Anderson, Intermetropolitan Migration: A Comparison of the Hypotheses of Zipf and Stouffer, *American Sociological Review,* Vol. XX (June, 1955), pp. 287–291.

the final population aged *n* years or more minus one-half of the net (positive or negative) migration, and (g) the final population *n* years of age and older minus all in-migrants plus those out-migrants who resided in the area initially.

Put in terms of the 1955–1960 illustration, the seven numbers are: (a) the April 1, 1960 population, (b) the April 1, 1955 population, (c) the October 1, 1957 population, (d) the population aged five years and older in 1960, (e) the mean of i (population aged five years or more in 1960) and ii (the 1955 population minus 1955–1960 deaths to 1955 residents), (f) the 1960 population five years of age or more minus half of the 1955–1960 net migration, and (g) the 1960 population aged five years or older minus 1955–1960 in-migrants plus 1955-1960 out-migrants who resided in the area in 1955.

These bases may be expressed in symbolic notation. The following terms identify the figures used in calculating the bases:

P_1 = the population of the area at time 1 (the beginning of the migration interval)

P_2 = the population of the area at time 2 (the end of the migration interval)

$P_{2,n+}$ = the population of the area at time 2 which is n years of age or older

D_{12} = deaths in the area during the interval between time 1 and time 2

$D_{12}|P_1$ = deaths in the area during the interval occurring to persons resident in the area at the beginning of the interval; i.e., deaths between time 1 and time 2 to persons resident in the area at time 1

I_{12} = in-migrants to the area during the interval between time 1 and time 2

O_{12} = out-migrants from the area during the interval between time 1 and time 2

$O_{12}|P_1$ = out-migrants from the area during the interval who resided in the area at the beginning of the interval; i.e., out-migrants between time 1 and time 2 of persons resident in the area at time 1

S_{12} = the population remaining in the area during the entire interval

In symbolic form, the ideal base population is: $S_{12} + O_{12}|P_1$. For 1955–60 migration, this becomes: $S_{1955\text{-}60} + O_{1955\text{-}60}|P_{1955}$. Unfortunately, data are rarely collected and tabulated in such a manner as to yield these two figures directly. Hence we resort to the manipulations of the table below.

Now let us discuss each proposal in turn. Proposal (a)'s defect is this: if an area has had a positive natural increase, base population (a) is too large by about half of the amount of the natural increase; hence the rates based on (a) as the denominator will be slightly too small. If the area has had an excess of deaths over births, population (a) is too small and the rates will be too large. If the area has increased through net migration, the final population is too great to the extent of half of the net migration; hence the rates will again be too small. If net migration is negative, the denominator is too small, making the rates too large. Last, and most distressing, if natural increase and net migration are both positive (or both negative), these errors are combined and would probably rise, for example, to seven or eight per cent for the fastest-growing states in the United States.

The second figure, (b), has the same disadvantages as (a) and therefore should be rejected quickly. Note that the errors inhering in (b) are opposite in direction to those of (a).

The mid-point population, while fully acceptable if one approves of tradition as a guide, is too crude an approximation to the four essential demographic qualities. Assume that during 1955–60, there were in a given country eight million births, six million deaths, and ten million migrants. Under these conditions, population (c) would be about four million too large because of including extraneous births and about three million too large because of including irrelevant deaths, a total error of about seven million. On the average, (c) would probably include about half of the out-migrants and half of the in-migrants, but the net effect would be zero (and therefore satisfactory) for very few political units. However, this is not to

Base	*Prose Definition*	*Computing Formula*	*1955–60 Example*
P_a	Final population	P_2	P_{1960}
P_b	Initial population	P_1	P_{1955}
P_c	Mid-period population	$P_1+\frac{P_2-P_1}{2}$	P_{1957}
P_d	Final population aged *n* years and older	$P_{2,n+}$	$P_{1960,5+}$
P_e	Arithmetic mean of i) final population aged *n* years and older and ii) initial population minus deaths to initial residents	$\frac{P_{2,n+}+(P_1-\left.D_{12}\right\|_{P_1})}{2}$	$\frac{P_{1960,5+}+(P_{1955}-\left.D_{1955\text{-}60}\right\|_{P_{1955}})}{2}$
P_f	Final population aged *n* years and older minus half the net migration	$P_{2,n+}-\frac{I_{12}-O_{12}}{2}$	$P_{1960,5+}-\frac{I_{1955\text{-}60}-O_{1955\text{-}60}}{2}$
P_g	Final population aged *n* years and older minus all in-migrants plus those out-migrants who resided in area initially	$P_{2,n+}-I_{12}+\left.O_{12}\right\|_{P_1}$	$P_{1960,5+}-I_{1955\text{-}60}+\left.O_{1955\text{-}60}\right\|_{P_{1955}}$

deny the appropriateness of the pre-eminence of the mid-point population as a base for fertility and mortality rates.

Population (d) must also be rejected. It is better than (c) because it does not include extraneous births or deaths, but it does not supply the two necessary migration qualities, i.e., numbers three and four of the four qualities enumerated in the last paragraph of Section II of this paper.

Proposal (e) is an average of two figures: the people in the area in 1960 who could have been there in 1955, and the people in the area in 1955 who could be there in 1960. The word "could" is inserted because the first condition assumes (ironically) no in-migration and the second condition assumes that we can separate deaths of in-migrants and of children born after April 1, 1955 from total deaths. A serious defect is that it would be very time-consuming and probably rather inaccurate to tabulate 1955–60 deaths by place of residence in 1955 for every area. If we use total 1955–60 deaths within the area, we have a figure inflated by deaths to in-migrants and by deaths to children born after 1955. Further, the assumption of zero in-migration is not realistic for most countries and cities.

Suggestion (f) is a modification of (d) to allow for migration. To illustrate its use, consider a city Q with a 1960 population of 100,000 aged five years or more and a net gain by migration of 2,000: base population (f) is then 100,000 minus half of 2,000 or 99,000. Again, if city Z has a population aged five years and over of 230,000 and a net migratory loss of 10,000, base population (f) is 230,000 plus 5,000, or 235,000. This figure is similar in principle to (e), but its determination is easier and more accurate. With regard to four essential qualities listed above, it does not include persons born after 1955 or persons dying before 1960, and it comes fairly close to the requirements concerning migration—under certain circumstances, errors attributable to the inclusion of in-migrants and non-1955-resident out-migrants may cancel each other out.[9]

[9] This cancellation occurs when the number of in-migrants is equal to the number of out-migrants who were residents of the area in 1955.

The last possibility, base (g), fits the four qualities exactly —by design, of course. It satisfies both the fertility and the mortality requirements (as do some of the other base populations), and it also meets the demands concerning both in- and out-migration (which none of the other bases do). Unfortunately, demographers are not always able to obtain the data needed to calculate (g). Although the age composition is usually known and the frequency of in-migrants is often obtainable, accurate tabulations of out-migrants by place of residence at the beginning of the migration period would be extremely difficult if not impossible to obtain in most instances. For example, for how many countries could we apportion 1955–1960 out-migrants into two categories: those resident in the area in 1955, and those not so residing?

IV. Conclusion

To summarize, there are four criteria for establishing the adequacy of a base population to be used in computing migration rates. Subsumed under the first criterion are four qualities or assumptions about the demographic variables: fertility, mortality, in-migration and out-migration.

Among seven bases analyzed in this article, five are unsatisfactory. Population (g) is the only one that fulfills completely these four stipulations of the first criterion, and it fails to comply with the other three criteria. Thus, no one population base is entirely suitable. The final choice in many parts of the world appears to be a compromise: population (f) is the closest approximation to the average exposed population that also meets the demands of the other three criteria: availability of data, accuracy of data, and ease of computation.

In practice, the choice is simple. If adequate data are available, the demographer should use the most logically rigorous base (g). Otherwise, he must use the second-best base (f).