Handling, Measuring, Estimating and Visualizing Migration Data in R

Guy J. Abel, James Raymer, Ellen Kraly

2023-05-19

Contents

Chapter 1

Introduction

This manual covers a range of methods for handling, measuring, estimating, and visualizing migration data in R. These methods are based on several authoritative sources, including the UN DESA manuals on *Methods of measuring internal migration* and *Preparing migration data for subnational population projections*, as well as the migration chapters of the IUSSP *Tools for Demographic Estimation*. Additionally, we will cover many more recent developments in the field. By the end of this manual, you will have a comprehensive understanding of the various methods available for working with migration data in R, and how to apply them to your own research.

To make the most of this manual, we assume that you have some knowledge of using R, especially the *tidyverse* set of packages. If you're not familiar with R or need a refresher, we recommend working your way through an online course before diving into this manual. Some good resources for learning R and the tidyverse include:

- *R for Data Science*, a comprehensive guide to data science in R, covering data import and cleaning, data visualization, and statistical modeling.
- *DataCamp*, an online learning platform that offers interactive courses on R programming and data science topics.
- *R Bootcamp*, a free online course that covers the basics of R programming and the tidyverse.
- *Tidyverse.org*, a website dedicated to the tidyverse packages with tutorials, articles, and other resources for learning and using them.
- *Swirl*, an interactive learning platform within R that teaches you how to use R in a hands-on manner.

These resources provide a comprehensive introduction to R programming and the tidyverse, which will be useful throughout this manual and in your future data analysis work.

The manual is organized into nine chapters, each covering a different aspect of handling, measuring, estimating, and visualizing migration data in R. Chapter 1 provides an introduction to migration data and concepts, while Chapter 2 focuses on handling migration data in R. Chapter 3 covers summary migration indices, while Chapter 4 is dedicated to estimating net migration. Chapter 5 focuses on describing and estimating migration age structure, and Chapter 6 covers describing bilateral migration data. Chapter 7 is dedicated to estimating bilateral migration, while Chapters 8 and 9 cover different methods for visualizing bilateral migration, including chord diagrams and Sankey plots. In each chapter, we provide code and data that will allow you to replicate the outputs we present, as well as exercises that allow you to practice the concepts covered in that chapter on your own. We also provide solutions to these exercises, so you can check your work and ensure that you have a firm grasp of the material before moving on to the next chapter. By following the examples and completing the exercises in this manual, you will gain a deep understanding of how to handle, measure, estimate, and visualize migration data in R.

Chapter 2

Migration Concepts

Migration is a complex phenomenon that involves a change in place of abode, or place of "usual" residence. As defined by the United Nations, migration refers to "the movement of a person or a group of persons, either across an international border or within a state" (**?**). It can refer to various demographic units, such as a person, a family, or a household. However, the definition of migration typically excludes certain groups, such as nomads who do not have a fixed place of residence, or seasonal migrants who live in two or more places during the course of a year.

Both a spatial (place) and temporal (change) dimension are required in the definition of migration. Spatially, migration involves moving from one place to another, whether it be across a border or within a state. Temporally, migration involves a change over time, such as taking up life in a new or different place. Unlike other demographic processes, migration is not a one-time event, but rather a process that can involve multiple moves over the course of a lifetime

2.1 Spatial

2.1.1 Usual Residence

Central to the spatial dimension of defining a migration event is the concept of place of residence, used to determine the migrants origin and destination locations. The *Principles and Recommendations for Population and Housing Censuses* (UN Statistics Division 2008: 102, para. 1.463) defines usual residence as follows:

"It is recommended that countries apply a threshold of 12 months when considering place of usual residence according to one of the following two criteria:

1. The place at which the person has lived continuously for most of the last

12 months (that is, for at least six months and one day), not including temporary absences for holidays or work assignments, or intends to live for at least six months

2. The place at which the person has lived continuously for at least the last 12 months, not including temporary absences for holidays or work assignments, or intends to live for at least 12 months."

The use of two alternative criteria leaves some area of ambiguity, where the subtle differences in each could have significant implications for the measurement of migration. Further, within either of these criteria they might be additional ambiguity. For example, persons on long work assignments might have intentions to stay for only a few months that might turn into many months.

When defining migration, typically no restrictions are placed on the distance involved in a relocation **?**, it could involve a move from one apartment to another in the same building or a move to another country. In the past, some researchers have distinguished between moves between local communities, such as cities or labor markets, and moves within local communities, often labeled as "migration" and "local mobility." However, many have argued that this distinction is problematic and that no spatial constraints on the definition of migration should be used. This is because such a distinction could be arbitrary and context-dependent.

When analyzing migration patterns, it is useful to have information on the distance involved in a relocation. If address information on points of origin and destination is available, it is possible to tabulate moves by the distance covered. However, in many countries without population registers, this is not possible, and it may be time-consuming and of little policy relevance.

Census or survey results are usually tabulated for the administrative or political units into which the country is divided. Therefore, migration is operationally defined as a change of residence from one civil division to another, and the volume of migration is then function of the size of areas chosen for compilation, where typically larger administrative units, such as a state or country, have a higher volume of migration compared to smaller units like cities or counties. Most countries typically have hierarchies in their administrative units and might provide migration data to reflect one or more of these geographies. Using larger geographic units can result in a loss of detail and accuracy in migration data. For instance, if a state is used as the unit of compilation, it may not capture migration patterns between different cities within the state, which could be important for local policy-making. On the other hand, using smaller geographic units such as census tracts may result in too much detail, making it difficult to draw broader conclusions about migration trends at the regional or national level.

2.1.2 Origin

When analyzing migration patterns, it is important to consider the different types of data that can be used. Most migration data can be categorized into different types based on the origin of migrants is defined. The two most commonly used types are migration event data, which is based on the previous place of residence, and migrant transition data, which is based on the place of residence a fixed number of years ago. Other types of migration data are occasionally collected that do not involve the definition of a migrants origin, including the duration at residence, number of moves over a given interval, and country of citizenship.

Lifetime migration data is another common type of migration data. It can be considered as a form of transition data, where the number of years changes based on the age of each individual. Migrant stock data is an aggregation over all persons' lifetime migration flow and is given at a specific point in time without an interval. The migration data literature often distinguishes between stock and flow data.

2.1.3 Migrant Transition Data

Migrant transition data are typically collected in national censuses, which identify migrants by comparing their place of usual residence at the time of enumeration (t) with that at a specified earlier date ($t-n$). This type of data provides information on the movements of migrants over a given time period, which is usually either 1 year (e.g. UK) or 5 years (e.g. USA). Some countries have time periods that correspond to the interval between the current and last census or significant time points in the country's history. Transition data have some limitation that become more prominent for migration measured using longer time periods. For example, migration transition data fail to identify multiple and return moves, which can lead to an underestimation of the true level of migration. Additionally, migrant transition data do not capture migrants who are born or who die during the measurement period, which can also impact the accuracy of the data.

Migrant transition data provide counts of migrants, where a migrant is defined as a person who has experienced one or more migrations during the specified prior period. It is important to note that persons who moved during the measurement interval and subsequently died before its end should technically be counted as migrants and their moves as migrations. However, in practice, such cases are usually excluded as information on migrants is usually obtained after the end of the interval and with reference to persons still living at that time. This exclusion can potentially lead to an underestimation of the true number of migrant transitions.

2.1.4 Migation Event Data

Event data records every move made by an individual, including multiple and return migrations, newborn moves, and moves immediately before death. Population registers typically collect these data and provide a more complete record of migration over time. However, the geographical units for which data are available are generally coarser, and registers often fail to capture information on within-region moves. Additionally, less information about the characteristics of migrants is usually available, and some groups may be omitted altogether, such as prisoners and military personnel.

There are important distinctions between the play (migration) and the actor (migrant). For a given migration interval, the number of migrants is rarely, if ever, as large as the number of migrations. Unless the interval is very short, such as a day or perhaps a week, some persons are certain to move more than once. The longer the migration interval, the more the count of migrants will understate the amount of migration. Conversely, the shorter the migration interval, the count of migrants will approach the number of migrations **?**.

2.2 Temporal

2.2.1 Migration Interval

Migration is a continuous process that occurs over time, and to study its incidence, data must be compiled with reference to specific periods of time. These time periods can be either definite or indefinite. Definite interval data is typically collected over fixed-term periods such as one year, five years, ten years, or the intercensal period. Indefinate interval data such as lifetime migration measures or data based on place of last residence lack a definite time reference as age or time at the current residence varies by each individual migrant.

The comparability of migration data with different definite time intervals can be prohibitively complicated. Commonly described as the one-year five-year problem, observed migration data consistently shows the number of migrants recorded over a five-year interval is far less than five times the number recorded over a one-year interval. In addition, the ratio of migrants between a five-year period and a one-year period is not constant, where variations occur depending on multiple factors such as the intensity and type of migration both over time and in each origin and destination. Consequently, there is no straightforward algebraic solution to comparing one-year and five-year migration probabilities **?**.

2.3 Migration Measures

Migration measures are used to quantify the magnitude and direction of population movements between places or regions. These measures can provide impor-

tant insights into the demographic and social dynamics of populations. There are several different types of migration measures that are commonly used in research and policy analysis, each with its own strengths and limitations.

One of the most common migration measures is the migration rate, which is defined as the number of migrants divided by the total population at risk. The migration rate can be calculated for different migration types discussed above, such as one year or five years, and can be used to compare migration across different places or regions.

Other migration measures include the count of the number of migrants, the migration intensity, which is the number of migrants per unit of population, and the migration propensity, which is the proportion of the population that have migrated. These measures can be useful for identifying patterns in migration behavior, such as the prevalence of long-distance migration or the likelihood of migration among certain demographic groups.

It is important to note that migration measures can be affected by data quality issues, such as underreporting of migrants or errors in place of residence information. Additionally, different measures may be more appropriate for different research questions or policy applications. For example, the migration rate may be more useful for understanding the overall magnitude of migration in the population, while the count of migrants can provide a basic understanding of the scale of migration patterns over time and between different spatial units.

Migration measures can be defined at different levels of detail, ranging from region-to-region measures, to region totals, to system totals or index measures. Region-to-region measures capture the flow of migrants between two specific regions, while region totals capture the total number of migrants coming in or going out of a specific region. System totals, or index measures, provide an overall picture of migration within a given system or country, which we will discuss in more detail in the next chapter.

2.3.1 Region-to-region

Region-to-region migration measures also known as *bilateral* migration, migration *streams* or *origin-destination* migration, refer to a migration measure that cross-classified by region of origin and region of destination, forming a matrix of $n \times (n-1)$ streams along each origin-destination combination, where n represents the number of regions. The set of region-to-region migration measures can be represented by m_{ij}, where the sub-scripts i and j represent the same set of regions for each origin-destination combination. The set of bilateral migration flows provide a basis to asses the compartive volumnes and directions of migration between a set of regions.

The *gross interchange* represents the total number of migrants moving between a particular pair of regions, i.e. $m_{ij} + m_{ji}$. The *net migration steam* or *bilateral net migration* represents the difference in migration between a pair of region

i.e. $m_{ij} - m_{ji}$. For a pair of streams that are of unequal size, where the net migration stream is not close to zero, there exists a *dominant stream* which is far large than the *reverse* or *counter* stream in the opposite direction.

2.3.2 Region Totals

Every migration event can be considered an out-migration in relation to the region of origin and an in-migration in relation to the region of destination. When migration events involve changes of countries, migration events are typically described as emigration and immigration, rather than out-migration and in-migration. Totals on in- or out-migration for each region are typically used to evaluate the volume of migration to or from a particular set of regions. In some countries, data is collected or aggregated without reference to the place of origin for in-migration totals or destination for out-migration totals. Consequently the migration totals provide the most detailed measure of regional migration but with little information on the direction of the migration flows between each region. A summary of the common terms for migration totals are shown in Table 1. The in-migration (or immigration) totals can be represented by replacing the origin i index with a +; m_{+j}. Similarly, the out-migration (or emigration) totals can be represented by replacing the destination j index with +; m_{+j}.

Scale	Area	Event Term	Migrant Term
Internal	Origin	out-migration	out-migrant
	Destination	in-migration	in-migrant
International	Origin	emigration	emigrant
	Destination	immigration	immigrant

The sum of the in-migration and out-migration totals ($m_{i+} + m_{+j}$) provides the *turnover* of each region. Net migration totals provides a balance of movements in opposing directions from the difference between in-migration and out-migration ($m_{+j} - m_{i+}$). Net migration measures are more typically obtained via demographic accounting, as a residual from the differences in population change, births and deaths over a period in each region. As this calculation does not require expensive migration data collection systems, net migration measures are one of the most common forms of migration measures. However, net migration measures have a number of notable drawbacks, as highlighted by **?**. In particular, net migration does not enumerate migrants themselves, but instead follows a residual of in-migrants and out-migrants. Consequently, the dynamics related to the observed migration patterns can be missed. For example, an net migration of -100 might involve a region receiving no in-migrants and sending 100 out-migrants or receiving 1,000,000 migrants and sending 1,000,100 out-migrants. Further migration dynamics are also missed when looking at net migration rates (discussed in the next section) and regularities in age profiles of

migration (discussed in Chapter X) are often preculded when using age-specicifc net migration measures.

2.3.3 Rate measures

Migration rates are important indicators for understanding the dynamics of population movement. Out-migration or emigration rates are calculated by dividing the number of out-migrants or emigrants during a specific period by the population exposed to the likelihood of migration. This is represented by the formula:

$$e^{[t,t+1]} = \frac{E^{[t,t+1]}}{P}k$$

Here, $e^{[t,t+1]}$ represents the out or emigration rate, E is the number of out-migrants or emigrants during the period, P is the population exposed to the likelihood of migration, and k is a constant, often set as 1000. The exposure population can be the population at the mid-interval, assuming migration is evenly distributed, or the population at the start or end of the interval if migration has a negligible effect on population change. Additionally, out-migration rates can be further decomposed by subsets of the population, such as age or sex:

$$e_i^{[t,t+1]} = \frac{E_i^{[t,t+1]}}{P_i}k$$

On the other hand, in-migration or immigration rates are calculated by dividing the number of in-migrants or immigrants by the population not exposed to the risk of migrating into the region. The formula for in-migration rate is:

$$i^{[t,t+1]} = \frac{I^{[t,t+1]}}{P}k$$

Similarly, net migration rates are calculated by dividing the net migration (difference between in-migration and out-migration) by the population not exposed to migration risk:

$$m^{[t,t+1]} = \frac{M^{[t,t+1]}}{P}k$$

It is worth noting that in-migration and net migration rates are different from other demographic rates because they use the resident population (population not exposed to risk) in the denominator. This approach satisfies the needs of the demographic balancing equation, as rates of gain and loss are measured relative to the same population. The demographic balancing equation is expressed as:

$$P^{t+1} = P^t \left(1 + b^{[t,t+1]} - d^{[t,t+1]} + i^{[t,t+1]} - e^{[t,t+1]}\right)$$

where P^{t+1} is the population at the next time point, $b^{[t,t+1]}$ and $d^{[t,t+1]}$ represent births and deaths during the period, and $i^{[t,t+1]}$ and $e^{[t,t+1]}$ denote in-migration and out-migration rates. Net migration ($M^{[t,t+1]}$) can be substituted with the difference between in-migration and out-migration ($I^{[t,t+1]} - O^{[t,t+1]}$). The equation can be simplified as:

$$P^{t+1} = P^t \left(1 + b^{[t,t+1]} - d^{[t,t+1]} + i^{[t,t+1]} - o^{[t,t+1]}\right)$$

This formulation allows for the analysis of population change considering the effects of births, deaths, in-migration, and out-migration.

2.4 References

Bibliography