A probabilistic framework for harmonization of migration statistics

Abstract. Inadequate and inconsistent data are a common and persistent problem in the field of migration. Deficiencies in migration statistics may be tackled using modelling techniques, something that has recently been recognized by European Union (EU) policymakers. The new Regulation on Community statistics on migration and international protection, which obliges countries to supply harmonized statistics, provides the possibility of using estimation methods to adapt statistics based on national definitions to comply with the required one-year duration of stay definition. The main objective of this chapter is to provide a theoretical probabilistic framework for capturing the various migration flow statistics that are available. It is a crucial step towards gaining a better understanding of the data and consequently harmonizing it. Different migration measures represent the same continuous data-generating process. They differ according to how the data happened to be collected and how the statistics happened to be produced. We introduce the key concepts of migration statistics using a simple duration model, namely an exponential distribution. While more complex models can better reflect the reality, they do not fundamentally modify the framework presented. The main focus is placed on the time criterion used in migration definition. This refers to duration of stay following relocation, which different countries specify very differently and which constitutes the main source of discrepancies in the operationalization of a migration concept in the EU member states.
3.1. Introduction

Data on international migration are lacking in terms of quality and cross-country comparability, which severely constrains analysis of migration patterns and their demographic, economic and social implications. The international migration debate in Europe and the European migration policy that is being implemented require, without doubt, high quality migration statistics that can be compared internationally. In August 2007, the new Regulation of the European Parliament and of the Council on Community statistics on migration and international protection entered into force (European Commission, 2007). The Regulation establishes a legal basis for the collection and compilation of migration statistics. It focuses on the comparability of statistical outputs and obliges member states, starting from the reference year 2009, to provide migration statistics that comply with a harmonized definition. The Regulation provides for the possibility of using statistical estimation methods to adapt statistics based on national definitions so that they comply with the harmonized definition. This emphasizes the importance of investigating such methods.

The purpose of this chapter is to present a probabilistic framework that is able to accommodate different definitions of migration and that may be used to convert different types of migration data into migration statistics with a harmonized definition. We intend to show that migration modelling is an effective approach to the harmonization of migration statistics. Currently, there is a considerable variability in the migration definitions applied by the countries of Europe. It results from the complexity of the migration process and the different national practices for measuring it. The key problem with defining migration stems from the fact that individual movements are situated in a time continuum. Spatial population movements include travel, commuting and migration. Migration is generally defined as a change of residence (address). However, the vagueness of residence and the coexistence of different types of residence (e.g. actual, usual and legal residence; temporary and permanent residence) lead to different conceptualizations of migration itself. An individual’s place of residence is usually determined by a duration-of-stay criterion, e.g. three months, one year or ‘permanent’. As a result, migration is a change of place of residence for at least three months, one year or ‘for good’ respectively (for details on migration flow statistics in the EU-25 see Kupiszewska and Nowok, 2008; Nowok et al., 2006). The duration of stay may be intended or actual. The intended duration of stay is based on a person’s intentions and these are usually revised over time as circumstances change. Consequently, it is possible that they differ from the actual length of stay.

A final operational definition of migration is very often a compromise between the concept of migration and available data sources. This increases the variability of possible measures. Courgeau (1973) introduced a crucial distinction between migrations and migrants. Essentially, migration count should refer to the number of moves and migrant count to the number of persons who move at least once during a reference period. Nonethe-
The number of migrants is often approximated through a typical census question about a place of residence at a previous date. Moreover, the migration definition may vary across subpopulations such as nationals and foreigners, for example. It may also be different for immigration and emigration, and it may change over time. There are numerous studies that discuss conceptual and measurement issues relating to migration, e.g. Bell et al. (2002), Bilsborrow et al. (1997), Poulain (1999; 2001), Poulain et al. (2006), United Nations (2002), Willekens (1982; 1985) and Zlotnik (1987).

The need to analyze migration patterns across time and countries has motivated the development of modelling techniques for overcoming the deficiencies present in migration statistics. Such attempts are, however, limited. Courgeau (1973) developed a model that relates the number of migrations to the census-based number of migrants. His method deals with multiple and return migrations. The hazard rates of migration are assumed to be constant and only part of the population can migrate again. Note that Courgeau’s model (1973) does not tackle the problems of migration definition itself. It was used mainly to study temporal trends in internal migration in France using census data for various geographical subdivisions (e.g. Baccaïni, 2007; Courgeau and Lelièvre, 2004). The model specification does not depend on spatial units that are analyzed, but the resulting parameter estimates are usually affected. The latter feature of the model also applies to the framework presented in this chapter.

A recently completed Eurostat project entitled MIMOSA – Migration Modelling for Statistical Analyses (http://mimosa.gedap.be/) – worked out a method for harmonizing international migration data available in Europe (De Beer et al., 2009). The authors use origin-destination specific flows as reported by sending and receiving countries to estimate a set of adjustment factors for both immigration and emigration figures that minimize the differences between the two available datasets. The correction factors are obtained using a constrained optimization procedure. In principle, this is the same approach to the harmonization of international migration data as that suggested by Poulain (1993) and later revised by Poulain and Dal (2008). A recent study by Abel (2009) provides a useful overview of the method and explores various alternative distance measures and constraint functions. Note that these methods do not provide answers about the linkage of one measure of migration to another. The values of the correction factors indicate the level of discrepancies between figures reported by different countries, but the definitional problems alone are not the cause of these differences.

This chapter focuses directly on migration definition. It approaches the migration process from a probabilistic perspective and views migration as a random event, i.e. an outcome of an underlying random process. By modelling the migration process, events and more particularly the distribution of events can be predicted. In studies of migration, a probabilistic approach is very natural and has been used for several decades (see e.g. Allison, 1985; Bijwaard, 2008; Constant and Zimmermann, 2003; 2007; Davies et al., 1982; Ginsberg, 1971; 1972; 1979a; 1979b; Pickles, 1983). The novelty of this study is...
that it applies probability theory to the harmonization of migration statistics. To tackle the issue properly, a distinction must be made between the migration process and the measurement process. Measuring is determining the magnitude or the characteristics of something. All measurements involve error but ideally errors remain within predefined limits. Unless the true process is known, measurement errors cannot be quantified. Hence, a few crucial questions have to be addressed before harmonization can be tackled. First, what is the true migration process? Second, how is migration measured? Third, what is the impact of the use of various measurements on the recorded level of migration flows? Finally, how can we obtain harmonized migration statistics from the available data? All these issues will be addressed in turn.

The chapter consists of five sections. Section 3.2 briefly presents the probabilistic model of migration, which is well documented in the literature. The basic parameter of the model is the instantaneous rate of relocation. This rate is referred to as relocation intensity or hazard rate of relocation. Section 3.3 reviews different measures of migration that are commonly used to produce migration statistics. In Section 3.4, the different migration measures are related to the basic parameters of the migration model. In other words, measures that result from different types of observation of migration are linked to the instantaneous rates of relocation, providing a powerful instrument for the harmonization of migration statistics. Section 3.5 concludes the chapter.

3.2. Migration process

There are two general approaches to modelling migration. The first is to model the data. A model is chosen that fits the data best, given a criterion of goodness of fit. In the second approach, an attempt is made to look behind the data and focus on the process itself. Model specification is of paramount importance here and the data are used to obtain the parameters of the model that is believed to describe the process accurately. The latter strategy, even though it may sometimes be speculative, should be given priority in the fields where very different measurements of the process are used. Migration is an obvious example of such a process. Thus, a migration process rather than migration data should be a point of departure.

We begin by assuming that migration is an unambiguously defined event that occurs at a specific point in time. Hereinafter this event is referred to as relocation, as distinct from operational definitions of the migration event that are used to produce migration statistics. In general terms, relocation is a change of residence (address). It may occur repeatedly for individuals at any point in time. A complete relocation history of an individual within a specific observation period is denoted here by \( \omega \). It may be presented in a compact way:
where $t_0$ is the onset of observation (beginning of the observed residence history) and $y_0$ the place of residence at that time, $t_n$ is the date of the $n$-th relocation and $y_n$ is the place of residence following the $n$-th relocation, $t_e$ denotes the end of observation and $y_e$ the place of residence at that time (Tuma and Hannan, 1984; Willekens, 1999). From this information we can infer where a person is living at every moment in the observation period.

From the perspective of stochastic processes expression (3.1) is a realization (sample path) of the underlying process. This relocation process may be described using counts (numbers of events in a given period of time) or waiting times (periods of time between successive events) (for a review of methods of analysis for repeated events see e.g. Cook and Lawless, 2002). In the context of migration statistics, aspects of both counts and waiting times are of particular relevance. We are interested in the total number of migrations, which are usually relocations with some conditions imposed on waiting times. Measures of migrations are discussed in detail in Section 3.3. The theory of counting processes (also referred to as arrival processes or point processes) therefore provides a useful general framework for the study of migration (Andersen et al., 1993). The counting process enables one to study the number and timing of events. It provides the possibility of making a straightforward connection between models for counts and duration models. Below we briefly describe a counting process and then the above-mentioned connection.

A counting process $\left\{ N(t) \mid t \geq 0 \right\}$ is a stochastic process which counts the number of events as they occur up to and including time $t$. The process has the properties that $N(0) = 0$, $N(t) < \infty$ with a probability of one and the sample paths of $N(t)$ are right-continuous and piecewise constant with jumps of size +1. The counting process is fully described by its random intensity process $\lambda(t)$ (for details on the concept of intensity see e.g. Blossfeld et al., 1989; Blossfeld and Rohwer, 2002; Klein and Moeschberger, 2003). For a short time interval $[t, t + dt)$, $\lambda(t) dt$ is the conditional probability of an event (relocation) in that interval, given all that has happened until just before $t$ (Aalen et al., 2008, pp. 26-27). Note that modelling recurrent events through their intensity functions is a very general and convenient approach. Let $T_n$ denote the arrival time of the $n$-th event. It is easy to observe that the time of $n$-th event is before or at $t$ if and only if the number of arrivals in $[0, t]$ is equal to $n$ or more. This reasoning gives the following relationship between waiting times and the number of events

$$T_n \leq t \iff N(t) \geq n.$$ (3.2)
Thus,

\[
P(N(t) = n) = P(N(t) \geq n) - P(N(t) \geq n + 1) = \]

\[
P(T_n \leq t) - P(T_{n+1} \leq t) = F_n(t) - F_{n+1}(t),
\]

(3.3)

where \( F_n(t) \) is the cumulative distribution function of \( T_n \). \( F_n(t) \) is also the \( n \)-fold convolution of the interarrival time distribution \( F(t) \) with itself, in other words the cumulative distribution function of the sum of \( n \) waiting times. Equation (3.3) provides the fundamental relationship between the distribution of waiting times and the distribution of counts.

A particularly simple duration model assumes that the hazard rate of relocations is constant, \( \lambda(t) = \lambda \). The time to event follows an exponential distribution. If interarrival times are independent and identically exponentially distributed, the counting process that results is a homogeneous Poisson process. Thus, a realization of a Poisson process can be seen as a sequence of realizations of independent exponentially distributed random durations whose lengths mark the occurrence of events in the process (Lancaster, 1990, p. 87). The number of events, \( N(t) \), in any fixed time interval from 0 to \( t \) follows a Poisson distribution with parameter \( \lambda t \):

\[
P(N(t) = n) = \frac{(\lambda t)^n \exp(-\lambda t)}{n!}, \quad n = 0, 1, 2, \ldots
\]

(3.4)

The parameter \( \lambda t \) is the expected number of events during the interval \((0, t)\). Note that probability functions of exponential and Poisson distributions apply for any interval of length \( t \), i.e. starting at any point on the time axis, not necessarily at the origin or event occurrence. Note that the probability that an individual does not experience an event during the interval is the survival function

\[
S(t) = \exp(-\lambda t)
\]

(3.5)

and the expected duration between successive relocations is equal to

\[
E[T_{n+1} - T_n] = \int_0^\infty S(t) \, dt = \int_0^\infty \exp(-\lambda t) \, dt = \frac{1}{\lambda}.
\]

(3.6)

The basic Poisson process may be generalized by allowing \( \lambda \) to differ between subpopulations and to vary in time. To take the differences between individuals into account we can introduce covariates to the model. Then the multiplicative hazards model due to Cox (1972), often called a proportional hazards model, is the most widely used one. An additional unobserved heterogeneity not captured by the observed characteristics may be represented by a random, discrete or continuous, variable. In modelling a positive continuous random effect the gamma distribution has a prominent role. In all the generalizations mentioned so far, however, we take the underlying assumption of exponentially distributed
interarrival times and the Poisson distribution for counts. A count data model with substantially higher flexibility than the Poisson model is obtained if we allow the intensity to vary not only between individuals but also with duration of stay. Distributions that capture duration dependence of the event occurrence include, inter alia, Weibull, Gompertz, gamma and lognormal distribution. Both Weibull and gamma distribution are generalizations of the exponential distribution and the resulting count data models nest the Poisson model. The specification of the count model that is consistent with an assumed waiting time distribution other than the exponential one is, however, not straightforward (see McShane et al., 2008 for Weibull distribution; and Winkelmann, 1995 for gamma distribution). In this study we use a Poisson process. This does not affect the basic idea of the framework presented. An extension of the model is necessary in order to capture better the complexities of either human behaviour or data collection systems that may function differently for nationals and foreigners, and for immigration and emigration.

3.3. Observation plans and measures

The relocation process is a continuous and recurrent phenomenon. To collect data generated by such a process, different observation plans, i.e. different schemes for collecting systematic information, can be used (Blossfeld and Rohwer, 2002; Tuma and Hannan, 1984). If we do not consider direction of relocation, the exact timings of all relocations experienced by each individual under study is the most complete information that can be available (compare with expression (3.1)). In practice, however, the collection of such relocation data is usually not feasible. For operational reasons, the migration event is defined in such a way that it can be practically measured. As a result, relocation processes are observed and measured in very different manners. It is of great importance, therefore, to understand the actual meaning of migration statistics in order to make a correct link with the underlying process. This section proposes a useful typology of existing migration data. The main data types are summarized in Table 3.1.

Recall first that the relocation history of an individual can be viewed from two different perspectives. In the first, the relocation history is described in terms of the events and their timing (event approach). In the second, the relocation history is described in terms of the places of residence at consecutive points in time (status approach). The intervals between the reference points can be of different lengths. Rajulton (2001) provides a direct connection between the event and status approaches by defining an event as a transition between statuses (states). Consider now a well-established distinction between migration data and migrant data. Essentially, migration denotes the act of moving (event) and migrant denotes the person performing the act (Courgeau, 1974). For a given reference period, a migrant is a person who moves at least once during this time interval. The
number of migrants is often estimated using a census or survey question about the place of residence at a previous date and is thus based on status data. As indicated by Courgeau (1973), this estimation is not satisfactory because return and non-surviving migrants are not enumerated. Nonetheless, in the migration literature the distinction described above between event data and status data (e.g. Ledent, 1980; Willekens, 1999) is usually treated as equivalent to the distinction between migration data and migrant data. Thus, in such an approach migrant denotes a person who moves at least once during a reference period and who lives in a different place at the end of the period than at the beginning. The event data and status data are also called movement data and transition data respectively (Rees and Willekens, 1986). As events are sometimes defined as transitions between statuses, for the sake of precision transition data can be called discrete transition data as opposed to direct transition data referring to movement data. In this study we distinguish three separate categories: migration data, migrant data (as defined by Courgeau, 1973) and discrete transition data (hereinafter referred to as transition data).

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Alternative names in the literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Conditional) migration</td>
<td>Event</td>
<td>Movement, direct transition</td>
</tr>
<tr>
<td>(Conditional) migrant</td>
<td>Person experiencing an event at least once during a reference period</td>
<td>-</td>
</tr>
<tr>
<td>Transition</td>
<td>Status of having a different place of residence at a specified date in the past</td>
<td>Migrant, discrete transition</td>
</tr>
</tbody>
</table>

We now introduce more specific data types that are particularly relevant for the harmonization of migration statistics. In official statistics the migration concept often involves a minimum duration of stay (actual or intended) to distinguish migration from other movements. Thus, migration is defined as a change in residence that is followed by a minimum duration of stay. The measurement of migration and migrants, which is conditional on a minimum duration of stay, leads to two data types that we call conditional migration data and conditional migrant data. The conditional migration data refer to migrations that are followed by a stay of specified duration, i.e. a person does not leave his or her new place of residence during that period. The conditional migrant data refer to migrants who experience at least one migration followed by a stay of a specified duration. As mentioned in the introduction, the duration may be intended or actual, where the former can be either shorter or longer than the latter. In this study we focus on actual duration assuming that all intentions are realized. The rationale behind the focus on conditional data types is the widespread use of an approach of this kind, especially in Europe. Note that data following a definition of a long-term migrant recommended by the United Nations (United Nations, 1998) falls into the category of conditional migrant data. This covers people who change their country of usual residence for a period of at least a year.
3.4. Indicators of migration process

As shown in the previous section, we received different results for the same underlying data-generating process depending on how the data happened to be collected and how the statistics happened to be produced. In this section we link empirical migration measures with an underlying relocation process. The connection is made through relocation intensity \( \lambda(t) \), which governs the process. For ease of exposition, we assume that members of a population migrate independently and that their migration experience may be described by the same Poisson process with the constant intensity \( \lambda \). The model was presented in Section 3.2. We start with the movement approach and consider the conditional migration and conditional migrant measures, and relationships between the two. Then, we present transition data and compare them with data produced using a movement approach.

Counting all relocations, without any restriction on the duration of stay in a destination place, leads to the expected number of \( \lambda t \) relocations in a time period of length \( t \) (hereinafter \( t \) without a subscript denotes the length of reference period). In practice, however, only selected relocations are counted as migrations. The concept of conditional migration, as described in Section 3.3, distinguishes migration from all relocations based on the minimum length of continuous stay that must follow change of place of residence. Thus, a person experiences a conditional migration when he or she changes place of residence and then does not do it again within a time interval of a fixed length of \( \hat{m}_t \). In other words, a person ‘survives’ time \( \hat{m}_t \) without any movement. Note that the requirement of continuity of stay is a simplifying assumption. In practice, some interruptions may occur, especially when a duration threshold of \( \hat{m}_t \) is relatively long. If the relocation rate is constant, the probability of being a stayer after \( \hat{m}_t \) is a survivor function of an exponential distribution or zero term in a Poisson distribution. Therefore, an expected number of conditional migrations with a duration threshold equal to \( \hat{m}_t \) experienced by an individual over a period of length \( t \) is derived from the Poisson distribution with a parameter corrected for survival of at least \( \hat{m}_t \):

\[
E \left[ N_{\hat{m}_t}(t) \right] = \sum_{n=0}^{\infty} \frac{(\lambda t \exp(-\hat{m}_t))^{n}}{n!} \exp(-\lambda t \exp(-\hat{m}_t)) = \lambda t \exp(-\hat{m}_t) \quad (3.7)
\]

The survivor function \( \exp(-\hat{m}_t) \) may be interpreted as the proportion of migrations that satisfy the duration-of-stay criterion. Thanks to a stochastic approach we know the chances of staying for various durations of \( \hat{m}_t \), even if the actual realizations take place beyond the reference period \( t \). In the special case when \( \hat{m}_t = 0 \) all relocations are counted. From (3.7) we discover an important relationship between counts of conditional migrations for two different durations of stay, \( \hat{m}_{t_1} \) and \( \hat{m}_{t_2} \):
HARMONIZATION BY SIMULATION

\[
E \left[ \frac{N_{t_m}(t)}{N_{t_{m_2}}(t)} \right] = \exp \left( -\lambda \left( t_m - t_{m_2} \right) \right),
\]  

(3.8)

The relationship depends on the relocation intensity but is independent of the length of the reference period \( t \). Below we present discrepancies between migration figures with different duration-of-stay criteria under various assumptions about relocation intensity. Since a one-year duration is recommended by the United Nations and required by the EU Regulation (European Commission, 2007; United Nations, 1998), we use it as a reference level. Thus, the values of the ratio (3.8) were calculated for different durations applied in the migration definition, \( t_m = t_m \in [0;5] \), relative to the UN definition, \( t_{m_2} = 1 \), and selected relocation intensity, \( \lambda \in (0;1] \). The choice of the considered values of \( t_m \) is determined by the lengths of duration criteria that are used in practice. Most often the duration threshold is equal to three months, six months or one year (Kupiszewska and Nowok, 2008). A threshold equal to zero refers to a migration definition with no duration criterion. Migration for at least five years may be seen as an approximation of a ‘permanent’ migration (Nowok, 2008). As regards the considered relocation intensities, the high values may be justified in the framework of a mover-stayer model. Only part of a population consists of potential migrants and the relocation intensity should therefore refer to these people. The results are presented in the left panel of Figure 3.1. For instance, if the migration intensity equals 0.2 (dotted line) and we count migrations for half a year, \( t_{m_2} = 0.5 \), instead of one year, we report figures that are higher by around 10 %. For the same migration rate of 0.2, counting migrations for five years, \( t_{m_2} = 5 \), results in an underestimation of the measure of migration by approximately 55 %. For the low levels of relocation intensities discrepancies between counts of migrations for different durations are relatively small. An increase in differences with a higher relocation rate results from the fact that with the increasing intensity a person relocates more often. In other words, durations between subsequent relocations become shorter and shorter and we observe multiple migrations for a short duration for the same individual and at the same time only a limited number of migrations for a longer duration. To get some idea of the discrepancies in actual migration data with different duration of stay criterion, compare figures on migration from Poland to Sweden in 1998-2007 produced by the two countries. This is equivalent to a comparison between the ‘permanent’ and one-year criterion used in Polish and Swedish data respectively. Depending on year, Poland reported numbers lower by 65-94 %. The disagreements may, however, also result from sources other than definitional ones, such as measurement errors.
Conditional migrant data show the same or lower discrepancies than conditional migration data. The reason for this is that migrant data do not count multiple migrations during the interval but only migrants who experienced at least one migration followed by a stay of specified duration. Note that, as described in Section 3.2, the concept of conditional migrant data differs from the concept of discrete transitions. Consider an individual who migrates twice during a reference period of one year. This person is counted as a conditional migrant if one of the relocations is followed by a stay of the duration in question. The person is included in the transition data if his or her place of residence at the end of the year differs from the place of residence at the beginning of the year. In other words, the second migration cannot be a return one. We calculated ratios analogous to (3.8) for conditional migrant data. Measures on migrants for different duration $t_{M_1}$ were compared with measures on migrants for one year, $t_{M_2} = 1$ ($M$ stands for migrants, to be distinguished from $m$ for migrations, which is of importance when both types of data are compared). They were, however, not derived analytically and the results of the microsimulation for annual data were therefore used instead. The resulting ratios for selected values of relocation intensity are shown in the right panel of Figure 3.1. The microsimulation was run in the R environment under the same assumptions about the relocation process as in case of conditional migration data.

Note that, unlike with conditional migration data, discrepancies between conditional migrant data for different durations depend on the length of the reference period $t$, which determines the possibility of multiple migrations of a specified duration. For instance, migration, neither for at least one year nor for five years, may not be experienced more than once within a one-year period. The annual numbers of conditional migrants for a one-year and a five-year stay, and consequently the ratio between the two, are exactly the same as the corresponding figures for conditional migrations. Within a three-year period,
multiple migrations are possible in the case of migration for one year but not for five years. As a result, the multiple migrations that are not included in statistics on migrants diminish the discrepancy between one-year and five-year conditional migrant data compared to conditional migration data. We focused our attention, however, on annual data because annual statistics are most common in practice. In fact, the impact of counting migrants instead of migrations on discrepancies between annual measures for different durations is of importance for a time criterion shorter than half a year. For longer durations the number of multiple migrants is negligible (see Figure 3.2).

![Figure 3.2](image)

**Figure 3.2** Conditional migrations per conditional migrant for the same duration \( t_m = t_M \), annual data

In principle, knowledge of the relocation rate enables us to recalculate counts of migrations or migrants for a specific duration (conditional migrations and conditional migrants respectively) into migrations or migrants for any other required duration. An example of the relationship between these types of annual measures for durations of up to one year and intensity \( \lambda = 0.2 \) is presented in Figure 3.3. The solid line represents a contour line of the value of one. For the corresponding pairs of duration thresholds \( t_m \) and \( t_M \) used in migration and migrant definitions respectively, the annual number of conditional migrations is equal to the annual number of conditional migrants. For instance, besides the obvious case of migrations and migrants for one year, the number of migrants for two months is approximately equal to the number of migrations for half a year. In other cases, if the data at our disposal refer to migrants for a specific duration and we would like to know the number of migrations for the same or different duration, we have to multiply our figure by the value indicated by the grey scale. For a relocation rate equal to 0.2, the discrepancy between the narrowest and the broadest measure within a one-year duration limit, namely the number of conditional migrants for one year, \( t_M = 1 \), and the number of all (non-conditional) migrations, \( t_m = 0 \), respectively, equals 22% (upper left corner of
This means that, during a period of one year, the number of migrations without any duration-of-stay restriction is 22% greater than the number of migrants under the one-year duration of stay criterion. If we raise the hazard rate from 0.2 to 0.4, the difference increases to 50%. Thus, for conditional measures of a duration of up to one year, which are usually used in practice, we should not expect differences greater than 50%. Nonetheless, if the widest measure is the conditional migrants for five years, which may approximate the measure of permanent migrants applied, for example, by some former state socialist countries, the difference increases to 172% for intensity $\lambda = 0.2$. For a migration rate equal to 0.4 the number of migrants for five years amount to less than 14% of the number of migrations without any duration of stay restriction. This percentage decreases rapidly with the increasing intensity, for instance, it amounts to 2% for $\lambda = 0.8$, but such a high international migration rate is vastly unrealistic.

![Figure 3.3 Ratio of conditional migrations to conditional migrants, for various durations up to one year and intensity $\lambda = 0.2$; solid line is a contour line of value one; dashed line is a line of equality of $t_m$ and $t_M$](image)

It is noteworthy that due to the distinction between migration and migrant measures, data with a longer duration-of-stay condition may be larger than data with a shorter one. In Figure 3.3, the area between the solid line (a contour line with a value of one) and the dashed line (a line of equality of duration condition in migration and migrant definition) includes combinations of lengths of duration threshold used in migration and migrant definition for which conditional migration numbers are greater than conditional migrant numbers, despite a longer duration criterion being used in the former case. For example, data on migrations for three months are larger by about 5% than data on migrants for one month. The number of combinations of duration thresholds for which the aforementioned relationship holds increases slightly with declining relocation intensity. At the same time, the lower the hazard rate of relocation, the lower the differences between the considered
measures. For a relocation intensity equal to 0.2 and 0.1 the discrepancies are smaller than 9 % and 5 % respectively.

So far we have considered conditional migration and conditional migrant measures, both of which are based on a movement approach. These data types are predominant in European statistical practice. Most of the official annual statistics on international migration flows produced in Europe represent one of these data types. Now we will consider a transition approach, i.e. direct transition measures that are based on the comparison of a person’s usual place of residence at two consecutive points in time. The data on international migration cover all individuals whose current place of usual residence is a country different from the one at a particular date in the past. The reference date is usually specified as one year or five years prior to enumeration. Such data are collected in many countries in a census or household survey, even if they are not used as a source of official statistics on international migration flows. Note that most of the few existing studies that address the issue of the relationship between different migration measures concentrate on this type of data derived for time intervals of various lengths, for example one and five-year periods (see Kitsul and Philipov, 1981; Liaw, 1984; Long and Boertlein, 1990; Rees, 1977; Rogers et al., 2003; Rogerson, 1990). We will first deal briefly with this type of comparability and look at the numbers of transitions for intervals of different length. Then we will compare transitions with conditional migrations.

Consider a simplified case when individuals relocate between two areas that form a closed system, with equal and constant intensity and relocations that occur independently of each other (some generalizations are amenable to calculations using matrix algebra). The chance $p$ of making a transition over a time interval, $t$, is equal to the chance of an odd number of relocations in this interval (compare Keyfitz, 1980):

$$p = \lambda t \exp(-\lambda t) + \frac{(\lambda t)^3 \exp(-\lambda t)}{3!} + \ldots = \frac{1 - \exp(-2\lambda t)}{2}. \quad (3.9)$$

When an individual relocates an even number of times between two areas he or she is in the same area at the beginning and end of the reference interval. This person does not contribute to the number of transitions and the total number of transitions does not increase linearly with time, as is the case for relocations. Nonetheless, for low relocation intensities the increase in transitions with the increasing length of reference interval is approximately linear (see Figure 3.4).
The relationship between numbers of transitions $N_p$ over time intervals of different lengths denoted by $t_{p_1}$ and $t_{p_2}$ is, based on expression (3.9), as follows

$$\frac{E[N_p(t_{p_1})]}{E[N_p(t_{p_2})]} = \frac{1 - \exp(-2\lambda t_{p_1})}{1 - \exp(-2\lambda t_{p_2})}.$$  

(3.10)

Figure 3.5 shows the ratio of transitions over few-years intervals to transitions over one year, depending on the level of relocation rate. The general decline in discrepancies between measures with a higher intensity results from the fact that an increase in hazard rate raises the chance of primary migration in short periods of time and repeat migrations in longer ones. The extreme values of rates for which different measures are hardly distinguishable are, however, presumably only theoretical. Consider transitions over a five-year interval compared with transitions over one year. Empirical five-year to one-year ratios reported in the literature for internal migration take on values of between two and four (Long and Boertlein, 1990; Rees, 1977; Rogers et al., 2003). They correspond to relocation intensity $\lambda$ between 0.06 and 0.33. Since internal migration is more prevalent than international migration we can expect that values of five-year to one-year ratios that are greater than four (hazard rate lower than 0.06) are quite realistic for international migration.
Figure 3.5 Ratio of transitions over an interval of different lengths to transitions over one year.

Under the simplifying assumptions stated above we can derive a relationship between transitions over intervals of a different length and conditional migrations for various durations of stay. We considered only the case when transitions and migrations are observed in intervals of the same length of $t$, i.e. when the reference period for conditional migrations number is equal to the interval over which we count the number of transitions. The length of duration criterion $t_m$ used in the migration definition may vary. For example, we compare the number of migrations that take place during a one-year reference period, $t=1$, and that are followed by at least a half-year stay, $t_m = 0.5$, with the number of people whose places of residence at the beginning and end of this reference year, $t=1$, differ. From (3.7) and (3.10) we obtain

$$\frac{E[N_{ho}(t)]}{E[N_{t_m}(t)]} = \exp(\lambda t_m)(1-\exp(-2\lambda t)) \frac{1}{2\lambda t}, \quad (3.11)$$

which enables us to go from events that occur during time $t$ and are followed by stays of various lengths of $t_m$ to transitions over periods of length $t$. For instance, if we know the annual number of migrations that are followed by at least a half-year stay and would like to obtain the number of transitions over the year, the figure has to be decreased by about 9%. Now consider the interesting case of discrepancies between the measure of international migration flows recommended by the United Nations for annual statistics and the measure of transitions over one year included in the census recommendations. For low relocation intensities the differences between these measures are negligible – for migration rates lower than 0.25 the differences are smaller than 1% (see solid and dashed lines in Figure 3.6).
For higher hazard rates the number of transitions over a one-year interval is higher than the number of conditional migrations for a one-year stay. This may come as a surprise, because the transition approach ignores multiple and return migrations within a reference interval. In the case of an annual measure of conditional migration for one year, multiple and return migrations are not possible. What is more crucial here, however, is that in the simplest transition approach applied above, a no-duration criterion is imposed on the length of stay in a current and reference place of residence. In practice, transitions are usually counted only for the resident population present in the country and residence is determined by the length of time that a person stays in the country. For illustrative purposes, the impact of a restriction on the minimum duration of stay in a current place of residence and also in a place of residence of one year before is presented in Figure 3.6 (dotted and dash-dotted lines). The results were obtained using microsimulation. The minimum length of stay was assumed to be half a year and this refers to actual total duration, i.e. for the current residence it includes time already spent and time that will be spent there in the future. The two additional constraints on minimum duration of stay decrease the number of transitions to the level lower than the numbers of conditional migrations for a one-year stay. This emphasizes the necessity of a careful consideration of not only a migration definition but also a definition of a resident population when different migration data are compared.

3.5. Conclusions

The inconsistency of statistics on international migration poses a persistent challenge for a comparative analysis of the phenomenon. This study has illustrated how the theory of
stochastic processes may yield important insights for an understanding of different migration measures and relationships between them. All migration measures represent the same underlying process, and estimates of the parameters of this process may be used to compute different quantities of interest. The main focus was placed upon the time criterion used in the measure of migration to select migrations from all changes of country of residence. The time refers to the duration of stay following relocation, which is specified very differently in different countries and constitutes the main source of discrepancies in the operationalization of the concept of migration in the EU member states. Under the simplifying assumptions that lead to a homogenous Poisson model of migration, a straightforward relationship exists between migration measures used in common migration statistics and relocation intensity. The hazard rate of relocation determines the level of discrepancies between different measures. The Poisson model used for illustrative purposes in this study may not be robust enough to provide an accurate description of all actual migration processes. It may be considered as a point of departure for more general counting processes that account for relocation intensities that vary according to duration of stay and across population groups. Future research should, therefore, test the simplifying assumptions about the underlying relocation process in a real-data situation. The straightforward approach is based on the likelihood of what is actually observed. Note, however, that individual relocation histories recorded in continuous time, which are best suited for estimates of relocation intensities, are often unavailable, and analysis has to rely on aggregate data. Moreover, in some cases the impact of definitional differences on migration numbers may be affected by accuracy or coverage problems.

References


