POPULATION, SPACE AND PLACE
Popul. Space Place **17**, 521–533 (2011)
Published online 28 July 2010 in Wiley Online Library
(wileyonlinelibrary.com) **DOI**: 10.1002/psp.624

A Probabilistic Framework for Harmonisation of Migration Statistics

Beata Nowok[1,*] and Frans Willekens[1]

[1]*Netherlands Interdisciplinary Demographic Institute, The Hague and Population Research Centre, Faculty of Spatial Sciences, University of Groningen, The Netherlands*

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, which has recently been recognised by the European Union (EU) policymakers. The new Regulation on Community statistics on international migration, which obliges countries to supply harmonised statistics, provides for the possibility of using estimation methods to adapt statistics based on national definitions to comply with the required 1-year duration of stay definition. The main objective of this paper is to provide a theoretical probabilistic framework to capture various migration flow statistics that are available. It is a crucial step towards better understanding and then harmonising the data. Different migration measures represent the same continuous data-generating process. They differ depending on 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 put on the time criterion used in the migration definition. It refers to the duration of stay following relocation, which is specified very differently among countries and constitutes the main source of discrepancies in operationalisation of the migration concept in the EU member states.

Accepted 18 June 2010

Keywords: international migration, migration concepts and measures, data harmonisation, duration of stay, counting process

INTRODUCTION

Data on international migration are lacking in 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 and internationally comparable migration statistics. 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 comparability of statistical outputs and obliges Member States to provide, starting from the reference year 2009, migration statistics that comply with a harmonised definition. The Regulation provides for the possibility of using statistical estimation methods to adapt statistics based on national definitions to comply with the harmonised definition, which emphasises the importance of investigating such methods.

The purpose of this paper is to present a probabilistic framework that is able to accommodate

*Correspondence to: Beata Nowok, Netherlands Interdisciplinary Demographic Institute, P.O. Box 11650, 2502 AR The Hague, The Netherlands. E-mail: beata.nowok@gmail.com

different definitions of migration and that may be applied to convert migration data of different types into migration statistics with a harmonised definition. We intend to show that migration modelling is an effective approach to the harmonisation of migration statistics. Currently, there is considerable variability in migration definitions applied by the countries of Europe. It results from the complexity of the migration process and different national practices to measure it. The essential 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 conceptualisations of migration itself. An individual's place of residence is usually determined based on the duration of stay criterion (e.g. 3 months, 1 year, or 'permanent'). As a result, migration is a change of place of residence for at least 3 months, 1 year, or 'for good', respectively (for details on migration flow statistics in the EU-25, see Nowok *et al.*, 2006; Kupiszewska and Nowok, 2008). The duration of stay may be intended or actual. Intended duration of stay is based on the person's intentions that are usually revised over time together with the changing circumstances, and eventually, 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 refers to the number of moves, and migrant count refers to the number of persons who move at least once during a reference period. Nonetheless, the number of migrants is often approximated through a typical census question about a place of residence at a previous date. Moreover, note that migration definition may vary across subpopulations such as nationals and foreigners, for example. It may be different for immigration and emigration, and it may change over time. There are numerous studies that discuss conceptual and measurement issues of migration (e.g. Willekens, 1982, 1985; Zlotnik, 1987; Bilsborrow *et al.*, 1997; Poulain, 1999, 2001; Bell *et al.*, 2002; United Nations, 2002; Poulain *et al.*, 2006).

The need to analyse migration patterns across time and countries has motivated the development of modelling techniques to overcome 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 the model of Courgeau (1973) does not tackle the problems of migration definition itself. It was used mainly to study temporal trends in internal migration in France based on census data for various geographical subdivisions (e.g. Courgeau and Lelièvre, 2004; Baccaïni, 2007). The model specification does not depend on spatial units that are analysed, but the resulting parameter estimates are usually affected. The latter feature of the model applies also to the framework presented in this paper.

A recently completed Eurostat project entitled MIMOSA – Migration Modelling for Statistical Analyses (http://mimosa.gedap.be/) worked out a method to harmonise international migration data available in Europe (De Beer *et al.*, 2009). The authors estimate, based on origin–destination specific flows as reported by sending and receiving countries, a set of adjustment factors for both immigration and emigration figures that minimise the differences between the two available data sets. The correction factors are obtained using a constrained optimisation procedure. In principle, this is the same approach to harmonisation of international migration data as suggested by Poulain (1993) and revised later 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 the 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 constitute only a part of these differences.

This paper focuses directly on migration definition. It approaches the migration process from

DOI: 10.1002/psp

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. Ginsberg, 1971, 1972, 1979a, b; Davies *et al.*, 1982; Pickles, 1983; Allison, 1985; Constant and Zimmermann, 2003, 2007; Bijwaard, 2008). The novelty of this study consists in applying probability theory to harmonisation of migration statistics. To properly tackle the issue, 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 harmonisation 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 harmonised migration statistics from the available data? All these issues are addressed in turn.

The paper consists of five sections. Migration process briefly presents the probabilistic model of migration, which is well documented in literature. The basic parameter of the model is the instantaneous rate of relocation. This rate is referred to as the relocation intensity or hazard rate of relocation. Observation plan and measures reviews different measures of migration, which are commonly used to produce migration statistics. In Indicators of migration process, 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 on migration are linked to the instantaneous rates of relocation, which provide a powerful instrument for the harmonisation of migration statistics. Conclusions concludes the paper.

MIGRATION PROCESS

There are two general approaches to modelling migration. The first is to model the data. A model is chosen, which fits the data best, given a criterion of goodness of fit. In the second approach, one attempts to look behind the data and focus on the process itself. Model specification is of paramount importance, and the data are used to obtain the parameters of the model that is believed to accurately describe the process. The latter strategy, although it may be sometimes 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.

Assume at first 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 ω. It may be presented in a compact way:

$$\omega[t_0, t_e] = \{t_0, y_0, t_1, y_1, \ldots, t_n, y_n, \ldots, t_e, y_e\}, \quad (1)$$

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 nth relocation, and y_n is the place of residence following the nth 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 in what place a person lives at every moment in the observation period.

From the perspective of stochastic processes, Equation (1) is a *realisation* (*sample path*) of the underlying process. This relocation process may be described through 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 basically relocations with some conditions imposed on waiting times. Measures of migrations are discussed in

DOI: 10.1002/psp

detail in the next section. The theory of *counting processes* (also referred to as *arrival processes* or *point processes*) provides, therefore, a useful general framework for the study of migration (Andersen *et al.*, 1993). The counting process enables one to study number and timing of events. It provides a possibility to make 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 $\{N(t)|t \geq 0\}$ is a stochastic process that 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 probability 1, 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 $[t, t + dt)$ (hereinafter, we use a square bracket to indicate inclusion of the interval endpoint, and a parenthesis for exclusion), $\lambda(t)dt$ is the conditional probability of an event (relocation) in that interval, given all that has happened until just before t (Aalen *et al.*, 2008: 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 nth event. It is easy to observe that the time of the nth 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:

$$T_n \leq t \Leftrightarrow N(t) \geq n. \tag{2}$$

Thus,

$$\begin{aligned} P(N(t) = n) &= P(N(t) \geq n) - P(N(t) \geq n+1) \\ &= P(T_n \leq t) - P(T_{n+1} \leq t) \\ &= F_n(t) - F_{n+1}(t), \end{aligned} \tag{3}$$

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) provides the fundamental relation 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, $\lambda(t) = \lambda$. Then, the time to the event follows an exponential distribution. If interarrival times are independent and identically exponentially distributed, the counting process that results is a homogenous Poisson process. Thus, a realisation of a Poisson process can be seen as a sequence of realisations of independent exponentially distributed random durations whose lengths mark the occurrence of events in the process (Lancaster, 1990: 87). The number of events $N(t)$ in any fixed time interval from 0 to t follows a Poisson distribution with parameter λt:

$$P(N(t) = n) = \frac{(\lambda t)^n \exp(-\lambda t)}{n!}, \quad n = 0, 1, 2, \ldots \tag{4}$$

The parameter λt 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 time axis), not necessarily being the origin or event occurrence. Note that the probability that an individual does not experience an event during the interval is the survival function $S(t) = \exp(-\lambda t)$, and the expected duration between successive relocations is equal to $\boldsymbol{E}[T_{n+1} - T_n] = \int_0^\infty S(t)\,dt = \int_0^\infty \exp(-\lambda t)\,dt = \frac{1}{\lambda}$. The basic Poisson process may be generalised by allowing λ to differ between subpopulations and to vary in time. To take the differences between individuals into account, we can introduce covariates in 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 generalisations mentioned thus far, however, the underlying assumption on exponentially distributed interarrival times and Poisson-distributed counts remains. 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 to vary with the duration of stay. Distributions that capture

DOI: 10.1002/psp

Table 1. Main types of migration data.

Type	Description	Alternative names in the literature
(Conditional) migration	*Event*	Movement, direct transition
(Conditional) migrant	Person experiencing an *event* at least once during a reference period	–
Transition	*Status* of having a different place of residence at a specified date in the past	Migrant, discrete transition

the duration dependence of the event occurrence include among others Weibull, Gompertz, gamma, and lognormal distribution. Both Weibull and gamma distribution are generalisations 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 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; nevertheless, this does not affect the basic idea of the framework presented. An extension of the model is necessary to better capture the complexities of either human behaviour or a data collection system that may function differently for nationals and foreigners, and for immigration and emigration.

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 (Tuma and Hannan, 1984; Blossfeld and Rohwer, 2002). If we do not consider the 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 Equation (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 ways. It is of great importance, therefore, to understand the actual meaning of migration statistics in order to make the correct link with the underlying process. This section proposes a useful typology of existing migration data. The main data types are summarised in Table 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, 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 through a census or survey question concerning the place of residence at a previous date, thus based on *status data*. As indicated by Courgeau (1973), this estimation is not satisfactory because returning and non-surviving migrants are not enumerated. Nonetheless, in the migration literature, the distinction between *event data* and *status data* described above (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 at the end of the period lives in a different place than at the beginning of the period. 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 a transition between statuses, to be more precise, transition data can be called *discrete transition data* as

DOI: 10.1002/psp

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*).

We now introduce more specific data types that are particularly relevant for the harmonisation of migration statistics. In official statistics, the migration concept often involves a minimum duration of stay (actual or intended) to distinguish migration from all 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, conditional on a minimum duration of stay, leads to two data types called by us *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 over that period). The *conditional migrant data* refer to migrants who experience at least one migration followed by a stay of 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 realised. 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) fall into the category of *conditional migrant data*. They cover persons who change their country of usual residence for a period of at least a year.

INDICATORS OF MIGRATION PROCESS

As presented in the previous section, for the same underlying data-generating process, we receive different results 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 constant intensity λ. The model was presented in Migration process. We start with movement approach and consider the *conditional migration and conditional migrant measures*, and the relationship between the two. Then, we present transition data and compare them with data produced based on movement approach.

Counting all relocations, without any restriction on the duration of stay in a destination place, leads to the expected number of λt 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*, described in the previous section, distinguishes migration from all relocations based on the minimum length of continuous stay that must follow a 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 t_m. In other words, a person 'survives' time t_m without any movement. Note that the requirement of a continuity of stay is a simplifying assumption. In practice, some interruptions may occur, especially when a duration threshold t_m is relatively long. When the relocation rate is constant, then the probability of being a stayer after t_m 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 t_m 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 t_m:

$$\begin{aligned} &\boldsymbol{E}\left[N_{t_m}(t)\right] = \\ &\sum_{n=0}^{\infty} n \frac{(\lambda t \exp(-\lambda t_m))^n \exp(-\lambda t \exp(-\lambda t_m))}{n!} \\ &= \lambda t \exp(-\lambda t_m). \end{aligned} \tag{5}$$

The survivor function $\exp(-\lambda t_m)$ may be interpreted as the proportion of migrations that satisfy the duration of the stay criterion. Thanks to the stochastic approach, we know the chances of staying for various durations t_m, even if the actual realisations take place beyond the reference period t. In the special case when $t_m = 0$, all relocations are counted. From Equation (5), we obtain an important relationship between counts of *conditional migrations* for different durations of stay, t_{m_1}, and t_{m_2}:

DOI: 10.1002/psp

$$\frac{\boldsymbol{E}\left[N_{t_{m_1}}(t)\right]}{\boldsymbol{E}\left[N_{t_{m_2}}(t)\right]} = \exp\left(-\lambda\left(t_{m_1} - t_{m_2}\right)\right). \qquad (6)$$

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 on relocation intensity. Because a 1-year duration is recommended by the United Nations (UN) and at the same time required by the European Union (EU) Regulation (United Nations, 1998; European Commission, 2007), we use it as a reference level. Thus, the values of Equation (6) are calculated for different durations applied in the migration definition, $t_{m_1} = t_m \in [0;5]$, 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 3 months, 6 months, or 1 year (Kupiszewska and Nowok, 2008). A threshold equal to 0 refers to a migration definition with no duration criterion. Migration for at least a 5-year stay may be seen as an approximation of a 'permanent' migration (Nowok, 2008). As regards considered relocation intensities, the high values may be justified in the framework of a mover–stayer model. Only a part of a population belongs to potential migrants, and the relocation intensity should refer to them. The results are presented in the left panel of Figure 1. For instance, if migration intensity equals 0.2 (dotted line) and we count migrations for half a year, $t_{m_1} = 0.5$, instead of 1 year, we report figures that are higher by around 10%. For the same migration rate of 0.2, counting migrations for 5 years, $t_{m_1} = 5$, 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 discrepancies occurs when 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 about the discrepancies in actual migration data with a 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 1-year criterion used in Polish and Swedish data, respectively. Depending on year, Poland reported numbers lower by 65–94%. The disagreements, however, may also be because of other reasons apart from definitional sources, such as measurement errors.

Conditional migrant data show the same or lower discrepancies than *conditional migration data*. The reason 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 Migration process, the concept of *conditional migrant data* differs from the concept of *discrete transitions*. Consider an individual who migrates two times during a reference period of 1 year. This person is counted as a conditional migrant if one of the relocations is

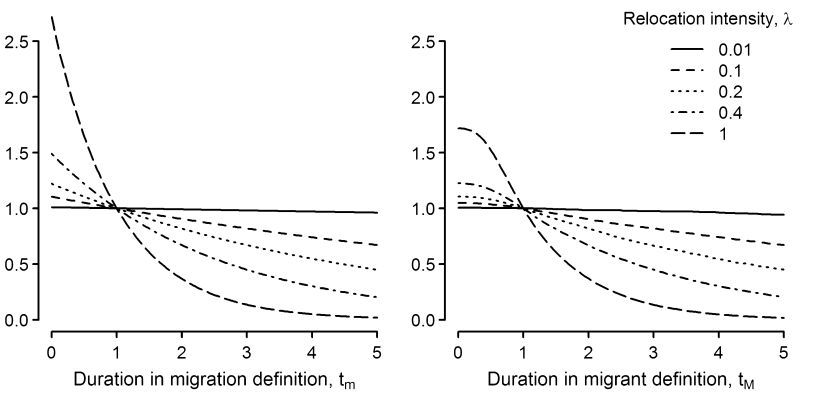

Figure 1. Ratio of conditional migration measures for various lengths of duration threshold to conditional measures for one year; left panel: conditional migrations, right panel: conditional migrants.

DOI: 10.1002/psp

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 Equation (6) 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 results of microsimulation for annual data were used instead. The resulting ratios for selected values of relocation intensity are shown in the right panel of Figure 1. Microsimulation was run in R environment under the same assumptions about the relocation process as in the case of *conditional migration data*. Readers interested in the use of microsimulation techniques for exploration of migration data may consult the forthcoming doctoral dissertation by Nowok.

Note that unlike for *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 for a specified duration. For instance, neither migration for at least 1 year nor for 5 years may be experienced more than once within a 1-year period. As a result, the ratio between the two is exactly the same as the corresponding figure for conditional migrations. Within a 3-year period, multiple migrations are possible in the case of migration for 1 year but not for 5 years. As a result, the multiple migrations that are not included in statistics on migrants diminish the discrepancy between 1-year and 5-year conditional migrant data compared with conditional migration data. We focus 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 time criterion shorter than half a year. For longer durations, the number of multiple migrants is negligible (see Fig. 2).

In principle, the knowledge of a relocation rate enables one 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 relations between these types of annual measures for durations up to 1 year and intensity $\lambda = 0.2$ is presented in Figure 3. The solid line represents a contour line of value 1. For the corresponding pairs of duration thresholds t_m and t_M used in migration and migrant definition, 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 1 year, the number of migrants for 2 months is equal approximately 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 gray scale. For a relocation rate equal to 0.2, within a 1-year duration limit, the discrepancy between the narrowest and the broadest measure, namely the number of conditional migrants for 1 year, $t_M = 1$, and the number of all (non-conditional) migrations, $t_m = 0$, respectively, equals 22% (upper left corner of Fig. 3). It means that, during a period of 1 year, the number of migrations without any duration of stay restriction is 22% larger than the number of migrants under the 1-year duration of stay criterion. If we raise the hazard rate from 0.2 to 0.4, the difference

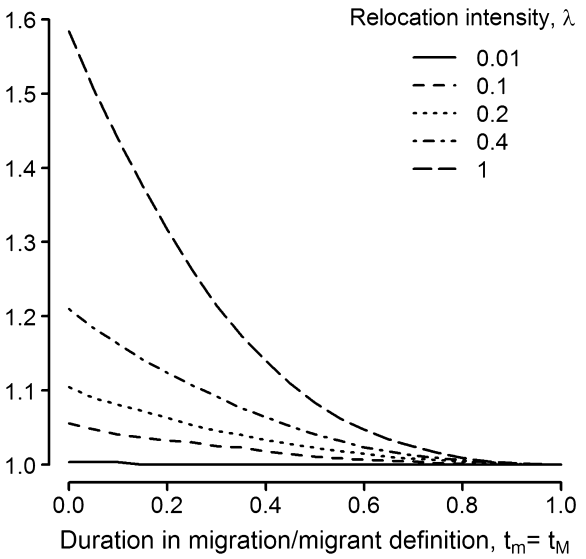

Figure 2. Conditional migrations per conditional migrant for the same duration $t_m = t_M$; annual data.

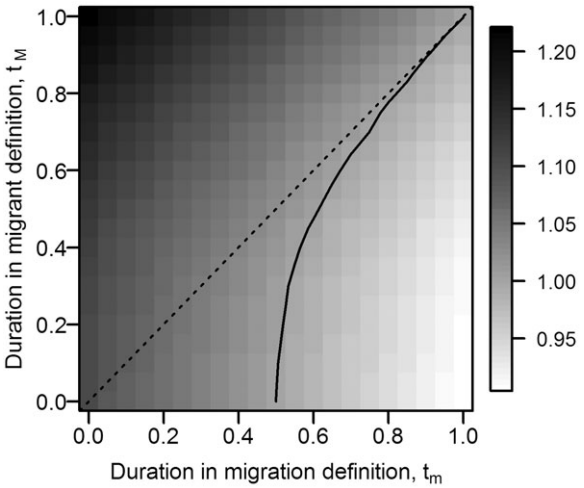

Figure 3. Ratio of conditional migrations to conditional migrants, for various durations up to one year and intensity $\lambda = 0.2$; solid line is a contour line of value one; dashed line is a line of equality of t_m and t_M.

increases to 50%. Thus, for conditional measures for a duration up to 1 year, which are usually used in practice, we should not expect differences larger than 50%. Nonetheless, if the widest measure is the conditional migrants for 5 years, which may approximate the measure of permanent migrants applied by, for example, 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 5 years amounts to less than 14% of the number of migrations without any duration of stay restriction. This percentage decreases rapidly with the increasing intensity (e.g. it amounts to 2% for $\lambda = 0.8$), but such a high international migration rate is vastly unrealistic nonetheless.

It is noteworthy that because of 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, the area between the solid line (a contour line of value one) and the dashed line (a line of equality of duration condition in the 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 used in the former case. For example, data on migrations for 3 months are larger than data on migrants for 1 month by about 5%. 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 relocation intensity equal to 0.2 and 0.1, the discrepancies are smaller than 9% and 5%, respectively.

Thus far, we have considered *conditional migration* and *conditional migrant* measures, which are based on the *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 them. Now, we consider the *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 in a country different from the one at a particular date in the past. The reference date is usually specified as 1 year or 5 years prior to enumeration. Such data are collected in many countries in census or household surveys, 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 relationships between different migration measures concentrate on this type of data derived for time intervals of various lengths (e.g. 1- and 5-year periods; see Rees, 1977; Kitsul and Philipov, 1981; Liaw, 1984; Long and Boertlein, 1990; Rogerson, 1990; Rogers *et al.*, 2003). We first deal briefly with this type of comparability and look at the numbers of transitions for intervals of different length. Then, we compare *transitions* with *conditional migrations*.

Consider a simplified case where individuals relocate between two areas that form a closed system with equal and constant intensity and relocations occur independently of each other (some generalisations 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):

$$p = \lambda t \exp(-\lambda t) + \frac{(\lambda t)^3 \exp(-\lambda t)}{3!} + \ldots = \frac{1 - \exp(-2\lambda t)}{2}. \quad (7)$$

DOI: 10.1002/psp

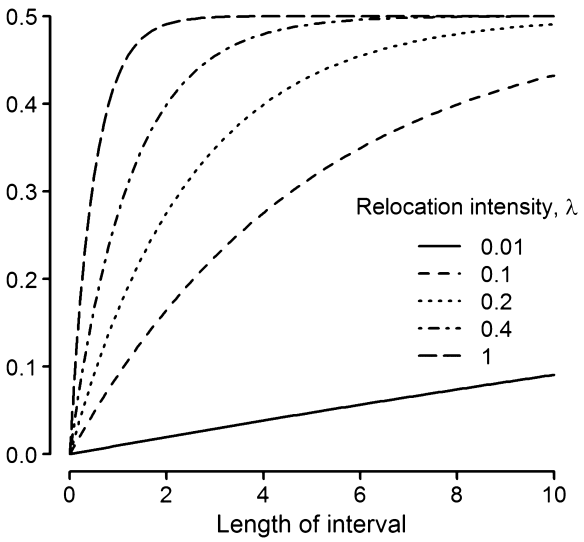

Figure 4. Expected number (per individual) of transitions over intervals of different lengths for selected intensities.

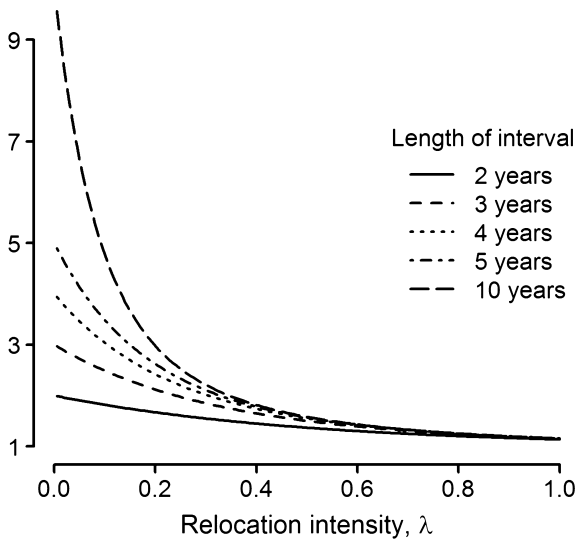

Figure 5. Ratio of transitions over an interval of different lengths to transitions over one year.

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 Fig. 4). The relation between numbers of transitions N_p over time intervals of different lengths denoted by t_{p_1} and t_{p_2} is, based on Equation (7), as follows:

$$\frac{\boldsymbol{E}\left[N_p(t_{p_1})\right]}{\boldsymbol{E}\left[N_p(t_{p_2})\right]} = \frac{1-\exp(-2\lambda t_{p_1})}{1-\exp(-2\lambda t_{p_2})} \tag{8}$$

Figure 5 shows the ratio of transitions over a few years' intervals to transitions over 1 year, depending on the level of relocation rate. The general decline in discrepancies between measures with higher intensity results from the fact that the 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 5-year interval compared with transitions over 1 year. Empirical 5-year to 1-year ratios reported in the literature for the internal migration take on values between 2 and 4 (Rees, 1977; Long and Boertlein, 1990; Rogers *et al.*, 2003). They correspond to relocation intensity λ between 0.06 and 0.33. Because internal migration is more prevalent than international, we can expect that values of 5-year to 1-year ratios greater than 4 (hazard rate lower than 0.06) are quite realistic for international migration nonetheless.

Under the simplifying assumptions stated above, we can derive a relationship between transitions over intervals of different length and conditional migrations for various durations of stay. We consider only the case when transitions and migrations are observed in intervals of the same length 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 migration definition may vary. For example, we compare the number of migrations that take place during a 1-year reference period, $t = 1$, and 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 of the end of this reference year, $t = 1$, differ. From Equations (5) and (8), we obtain

$$\frac{\boldsymbol{E}\left[N_p(t)\right]}{\boldsymbol{E}\left[N_{t_m}(t)\right]} = \frac{\exp(\lambda t_m)(1-\exp(-2\lambda t))}{2\lambda t}, \tag{9}$$

DOI: 10.1002/psp

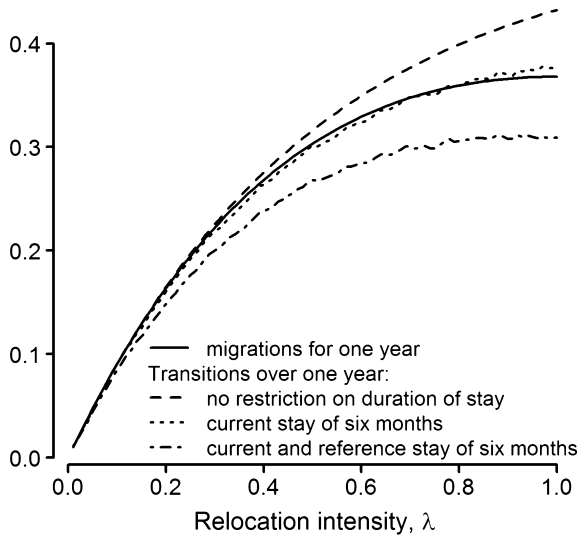

Figure 6. Expected number (per individual) of conditional migrations for one year and transitions over one year with and without restriction on minimum duration of residence.

which enables us to go from events that occur during time t and are followed by stays of various length 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 UN for annual statistics and the measure of transitions over 1 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 Fig. 6). For higher hazard rates, the number of transitions over a 1-year interval is higher than the number of conditional migrations for a 1-year stay. It may come as a surprise because the transition approach ignores multiple and return migrations within a reference interval. In the case of the annual measure on conditional migration for 1 year, multiple and return migrations are not possible nonetheless. But what is more crucial here is that in the simplest transition approach applied above, 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 a residence is determined based on the length of time that a person stays in the country. For illustrative purposes, the impact of restriction on the minimum duration of stay in a current place of residence and also in a place of residence occupied 1 year before is presented in Figure 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 it refers to actual total duration (i.e. for the current residence it includes time already spent and time that will be spent 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 1-year stay. This emphasises the necessity of a careful consideration of not only a migration definition, but also of a definition of a resident population, when different migration data are compared.

CONCLUSIONS

The inconsistency of statistics on international migration poses a persistent challenge for a comparative analysis of the phenomenon. This study illustrated how theory of stochastic processes may yield important insights into understanding of different migration measures and relationships between them. All migration measures represent the same underlying process, and estimates of its parameters may be used to compute diverse quantities of interest. The main focus was put on the time criterion used in the migration measure 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 among countries and constitutes the main source of discrepancies in the operationalisation of migration concept in the EU member states. Under the simplifying assumptions leading 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 in this study for illustration purposes may not be robust enough to give an accurate description of all actual migration processes. It may be treated as a point of departure for more general counting processes that account

DOI: 10.1002/psp

for relocation intensities that vary with 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 very often not available, and analysis has to rely on aggregate data. Moreover, in some cases, the impact of definitional differences on migration numbers may be disturbed by accuracy or coverage problems.

REFERENCES

Aalen OO, Borgan Ø, Gjessing HK. 2008. *Survival and Event History Analysis: a Process Point of View.* Springer: New York.

Abel GJ. 2009. *International Migration Flow Table Estimation*. PhD thesis, University of Southampton, School of Social Sciences.

Allison PD. 1985. Survival analysis of backward recurrence times. *Journal of the American Statistical Association* **80**: 315–322.

Andersen PK, Borgan Ø, Gill RD, Keiding N. 1993. *Statistical Models Based on Counting Processes.* Springer-Verlag: New York.

Baccaïni B. 2007. Inter-regional migration flows in France over the last fifty years. *Population-E* **62**: 139–156.

Bell M, Blake M, Boyle P, Duke-Williams O, Rees P, Stillwell J, Hugo G. 2002. Cross-national comparison of internal migration: issues and measures. *Journal of the Royal Statistical Society: Series A* **165**: 435–464. DOI: 10.1111/1467-985X.00247.

Bijwaard G. 2008. Immigrant migration dynamics model for The Netherlands. *Journal of Population Economics*. DOI: 10.1007/s00148-008-0228-1 [accessed 10 April 2010].

Bilsborrow RE, Hugo G, Oberai AS, Zlotnik H. 1997. *International Migration Statistics: Guidelines for Improving Data Collection Systems.* International Labour Office: Geneva.

Blossfeld HP, Hamerle A, Mayer KU. 1989. *Event History Analysis: Statistical Theory and Application in the Social Sciences.* Lawrence Erlbaum Associates: New Jersey.

Blossfeld HP, Rohwer G. 2002. *Techniques of Event History Modeling: New Approaches to Causal Analysis.* Lawrence Erlbaum Associates: New Jersey.

Constant A, Zimmermann K. 2003. The dynamics of repeat migration: a Markov chain analysis. *IZA Discussion Paper* **885**: 1–35.

Constant A, Zimmermann K. 2007. Circular migration: counts of exits and years away from the host country. *IZA Discussion Paper* **2999**: 1–23.

Cook RJ, Lawless JF. 2002. Analysis of repeated events. *Statistical Methods in Medical Research* **11**: 141–166.

Courgeau D. 1973. Migrants et migrations. *Population* **28**: 95–129 (Also in English: Migrants and migrations. *Population, Selected Papers* 1979, **3**: 1–35).

Courgeau D. 1974. Methodological aspects of the measurement of international migration. In *International Migration Review*, Tapinos G (ed.). Committee for International Coordination of National Research in Demography: Paris; 69–82.

Courgeau D, Lelièvre E. 2004. Estimation of French internal migration in the period 1990–1999 and comparison with earlier periods. *Population-E* **59**: 703–710.

Cox DR. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B* **34**: 187–220.

Davies RB, Crouchley R, Pickles AR. 1982. Modelling the evolution of heterogeneity in residential mobility. *Demography* **19**: 291–299.

De Beer J, Van der Erf R, Raymer J. 2009. *Estimates of OD matrix by broad group of citizenship, sex and age, 2002–2007. Report for the MIMOSA project*. Available at http://mimosa.gedap.be/Documents/Mimosa_2009b.pdf [accessed 10 April 2010].

European Commission. 2007. *Regulation (EC) No 862/2007 of the European Parliament and of the Council of 11 July 2007 on Community statistics on migration and international protection*. European Commission: Brussels. Available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:199:0023:0029:EN:PDF [accessed 10 April 2010].

Ginsberg RB. 1971. Semi-Markov processes and mobility. *Journal of Mathematical Sociology* **1**: 233–262.

Ginsberg RB. 1972. Critique of probabilistic models: Application of the semi-Markov model to migration. *Journal of Mathematical Sociology* **2**: 63–82.

Ginsberg RB. 1979a. Timing and duration effects in residence histories and other longitudinal data: I – stochastic and statistical models. *Regional Science and Urban Economics* **9**: 311–331.

Ginsberg RB. 1979b. Timing and duration effects in residence histories and other longitudinal data: II – studies of duration effects in Norway, 1965–1971. *Regional Science and Urban Economics* **9**: 369–392.

Keyfitz N. 1980. Multistate demography and its data: a comment. *Environment and Planning A* **12**: 615–622.

Kitsul P, Philipov D. 1981. The one-year/five-year migration problem. In *Advances in Multiregional Demography*, Rogers A (ed.). International Institute for Applied Systems Analysis: Research Report RR-81-6, Laxenburg: Austria; 1–33.

Klein JP, Moeschberger ML. 2003. *Survival Analysis: Techniques for Censored and Truncated Data.* Springer: New York.

Kupiszewska D, Nowok B. 2008. Comparability of statistics on international migration flows in the European Union. In *International Migration in Europe: Data, Models and Estimates*, Raymer J, Willekens F (eds). John Wiley & Sons, Ltd: Chichester; 41–71.

Lancaster T. 1990. *The Econometric Analysis of Transition Data.* Cambridge University Press: New York.

Ledent J. 1980. Multistate life tables: movement versus transition perspectives. *Environment and Planning A* **12**: 533–562.

Liaw KL. 1984. Interpolation of transition matrices by the variable power method. *Environment and Planning A* **16**: 917–925.

Long JF, Boertlein CG. 1990. Comparing migration measures having different intervals. U.S. Census Bureau: Washington. *Current Population Reports, Series P-23* **166**: 1–11.

McShane B, Adrian M, Bradlow ET, Fader PS. 2008. Count models based on Weibull interarrival times. *Journal of Business and Economic Statistics* **26**: 369–378. DOI: 10.1198/073500107000000278.

Nowok B. 2008. Evolution of international migration statistics in selected Central European countries. In *International Migration in Europe: Data, Models and Estimates*, Raymer J, Willekens F (eds). John Wiley & Sons, Ltd: Chichester; 73–87.

Nowok B, Kupiszewska D, Poulain M. 2006. Statistics on international migration flows. In *THESIM: Towards Harmonised European Statistics on International Migration*, Poulain M, Perrin N, Singleton A (eds). Presses Universitaires de Louvain: Louvain-la-Neuve; 203–231.

Pickles AR. 1983. The analysis of residence histories and other longitudinal panel data: a continuous time mixed Markov renewal model incorporating exogenous variables. *Regional Science and Urban Economics* **13**: 271–285.

Poulain M. 1993. Confrontation des Statistiques de migrations intra-européennes: Vers plus d'harmonisation? *European Journal of Population* **9**: 353–381.

Poulain M. 1999. *International migration within Europe: towards more complete and reliable data?* Paper presented at the *Joint ECE-Eurostat Work Session on Demographic Projections*, Perugia, Italy, May 1999.

Poulain M. 2001. *Is the measurement of international migration flows improving in Europe?* Paper presented at the *Joint ECE-Eurostat Work Session on Migration Statistics*, Geneva, 2001.

Poulain M, Dal L. 2008. *Estimation of flows within the intra-EU migration matrix. Report for the MIMOSA project*. Available at http://mimosa.gedap.be/Documents/Poulain_2008.pdf [accessed 10 April 2010].

Poulain M, Perrin N, Singleton A (eds). 2006. *THESIM: Towards Harmonised European Statistics on International Migration.* Presses Universitaires de Louvain: Louvain-la-Neuve.

Rajulton F. 2001. Analysis of life histories: a state space approach. *Canadian Studies in Population* **28**: 341–359.

Rees P. 1977. The measurement of migration, from census data and other sources. *Environment and Planning A* **9**: 247–272.

Rees P, Willekens F. 1986. Data and accounts. In *Migration and Settlement: a Multiregional Comparative Study*, Rogers A, Willekens F (eds). Reidel Press Dordrecht; 19–58.

Rogers A, Raymer J, Newbold KB. 2003. Reconciling and translating migration data collected over time intervals of differing widths. *Annals of Regional Science* **37**: 581–601.

Rogerson PA. 1990. Migration analysis using data with time intervals of differing widths. *Papers in Regional Science* **68**: 97–106.

Tuma NB, Hannan MT. 1984. *Social Dynamics: Models and Methods.* Academic Press: London.

United Nations. 1998. *Recommendations on Statistics of International Migration: Revision 1.* Statistical Papers, No. 58, Rev.1 Sales No. E.98.XVII.14: New York.

United Nations. 2002. *International Migration Report 2002.* United Nations Population Division, Department of Economic and Social Affairs: New York.

Willekens F. 1982. Identification and measurement of spatial population movements. In *A National Migration Survey, Manual X: Guidelines for Analysis*. United Nations ESCAP: New York; 74–97.

Willekens F. 1985. Comparability of migration data: Utopia or reality? In *Migrations Internes, Collecte des Données et Méthodes d'Analyse*, Poulain M (ed.). Cabay: Louvain-la-Neuve; 409–441.

Willekens F. 1999. Modeling approaches to the indirect estimation of migration flows: from entropy to EM. *Mathematical Population Studies* **7**: 239–278.

Winkelmann R. 1995. Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics* **13**: 467–474.

Zlotnik H. 1987. The concept of international migration as reflected in data collection systems. *International Migration Review* **21**: 925–946.