Towards harmonisation of migration statistics: a probabilistic perspective.

Beata Nowok[*]**,** *Population Research Centre, University of Groningen*
Frans Willekens, *Netherlands Interdisciplinary Demographic Institute*

1. Introduction

The international migration debate in Europe and the European migration policy that is being implemented require 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 harmonized definition. The Regulation provides for possibility of using statistical estimation methods to adapt statistics based on national definitions to comply with the harmonised definition. The applied methods must be, however, scientifically based and well-documented.

The purpose of this paper is to present a model of migration flows that is able to accommodate different definitions of migration and that may be applied to convert migration data of different type 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 considerably variability in migration definitions applied by the countries of Europe. It results from the complexity of migration process and different national practices to measure it. The essential problems with defining migration stem from the fact that individual movements are situated in time continuum. Spatial population movements include travel, commuting and migration. Migration is generally defined as a change of residence (address) that involves the crossing of an administrative boundary. However the vagueness of residence and coexistence of different types of residence (e.g. actual, usual and legal residence; temporary and permanent residence) lead to different conceptualization of migration itself. Individual's place of residence is usually determined based on a duration of stay criterion, e.g. 3 months, 1 year, or "permanent". In 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 Kupiszewska and Nowok forthcoming, and Nowok et al.

[*] Email: nowok@nidi.nl

2006). The duration of stay may be intended or actual. Intended duration of stay is based on person's intensions that are usually revised over time together with the changing circumstances and eventually they differ from the actual length of stay.

Different operationalizations of the migration concept increase the variability of possible measures. Numerous studies discuss conceptual and measurement issues, for instance, Bell et al. (2002), Bilsborrow et al. (1997), Poulain (1999, 2001), Poulain et al. (2006), United Nations (2002), Willekens (1982, 1985), and Zlotnik. Courgeau (1973, 1979) introduced a crucial distinction between migrations and migrants. That distinction and other differences in concepts and measures were later accommodated in data types and observational plans.

This paper approaches the migration process from a probabilistic perspective and view migration as a random event that it 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 already used in modelling migration process by many scholars (see e.g. Ginsberg 1979a, 1979b, 1972, 1971). The novelty of this study consists in applying probability theory to harmonization 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 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 to obtain harmonized migration statistics from the available data? All these issues are addressed in turn.

The paper consists of 5 sections. Section 2 briefly presents the probabilistic model of migration. The model is well-documented in the literature. The basic parameter of the model is the instantaneous rate of migration or migration intensity. Section 3 reviews different measurements of migration that are revealed in migration statistics published in Europe. In Section 4, the different migration measures are related to the basic parameters of the migration model. Linking different types of observation on migration to the instantaneous rates of migration that are consistent with a given model provides a powerful instrument for the harmonization of migration statistics. Section 5 concludes the paper.

2. Migration process

There are two general approaches to modelling processes. 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, 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, even though 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 is a point of departure and reference in this study.

Note that we do not address the direction of migration. In addition we assume that migration is an unambiguously defined event that occurs at a specific point in time (we disregard the fact that moving from one place to another takes time). The migration event is defined as a change of residence (address) involving the crossing of administrative boundary. It may occur repeatedly, thus migration is a recurrent event. In addition, it may take place at any point in time. Migration status is determined based on comparison of places of residence at different dates. The status of being a migrant is attached to persons whose current place of residence is located in a different area than it was at a prior date.

Consider the migration history of an individual. The sequence of migrations may be presented in a compact way:

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

where t_0 is the onset of observation (beginning of the residence history) and y_0 the place of residence at that time, t_n is the date of the *n*-th migration and y_n is the place of residence following the *n*-th migration, t_e the end of observation and y_e the place of residence at that time (Tuma and Hannan 1984, Willekens 1999). From (1) we can infer in what place a person lives at every moment in the observation period. Thus, it gives an opportunity to describe migration histories in terms of events and the times to events, or in terms of the place of residence at different points in time. The first approach is referred to as the event approach and the second as the status approach. Migration statuses may be derived for intervals of any specified length. A complete residence history (1) starts at birth and ends at death. In practice the limits are set by the observation period.

An observed residence history of a person $\omega[t_0, t_e]$ is a realization of a stochastic process. A stochastic process may yield many different realizations. In the theory of stochastic processes, a particular realization is called a *sample path*. The theory of *counting processes* (also referred to as *arrival processes* or *point processes*) provides a

general theoretical framework for the study of repeated events (or "*arrivals"*) such as migration (Andersen et al. 1993). The counting process enables to study number and timing of events. It makes connection between models for counts and duration models. A duration model describes the distribution of time-to-event or waiting time given the model and the instantaneous rates of transition. Models of counts describe the probability distribution of numbers of events given the model and the intensities. A counting process $\{N(t) | t \geq 0\}$ is a stochastic process which counts the number of events of interest by time *t*. The process has the properties that $N(0) = 0$, $N(t) < \infty$ with probability one, and the sample paths of $N(t)$ are right-continuous and piecewise constant with jumps of size +1. The distributions of time to event and number of events during a given interval are fully determined if the instantaneous rates are known. Let $\mu(t)$ denoted the instantaneous rate of migration at time t. For a short time interval $[t, t+dt)$, $\mu(t)dt$ is the conditional probability of an event (migration) in that interval given all that has happened until just before *t* (Klein and Moeschberger 2003, pp. 79-80). The parameter of the counting process model is the number of events per unit of time. It is denoted by $\lambda(t)$. Note that $\lambda(t) = \mu(t)$ provided $\mu(t)$ is constant in an interval of unit length. In this paper, we characterize processes in terms of $\lambda(t)$. Distributions that characterize the time to event in case of a non-repeatable event and the interval between events in case of repeatable events are the survival function $S(t)$, the cumulative distribution function $F(t)$, the hazard function $\lambda(t)$ and the density function $f(t)$. The probability distribution of the number of events during a given period $P(N(t))$ also characterizes the process.

The parameter $\lambda(t)$ varies with time *t* and the time dependence of λ may itself be modelled. Different time dependencies result in different duration models. A particularly simple duration model assumes that the transition rate is constant. Thus, the hazard function is

$$\lambda(t) = \lambda, t \geq 0, \qquad \lambda > 0. \tag{2}$$

If the transition rate is constant, the time to event follows an exponential distribution. The respective survivor, cumulative distribution and density functions of an exponential distribution are

$$S(t) = e^{-\lambda t}, \tag{3}$$

$$F(t) = 1 - e^{-\lambda t}, \tag{4}$$

$$f(t) = \lambda e^{-\lambda t}. \tag{5}$$

In case the event is repeatable and the event rate does not depend on the number of occurrences, the interarrival times are independent and identically exponentially distributed. The counting process that results is a homogenous Poisson process that is characterized by the Poisson distribution. 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, p. 87). The number of events $N(t)$ in any fixed time interval from 0 to *t* follows a Poisson distribution with parameter λt:

$$P(N(t)=n)=\frac{(\lambda t)^n e^{-\lambda t}}{n!}, \qquad n=0,1,2,... \tag{6}$$

The parameter λt is the expected number of events during the interval (0,t):

$$E[N(t)]=\lambda t\,. \tag{7}$$

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. This property is called lack of memory and the exponential distribution is the unique continuous distribution possessing this characteristic. Note that the probability that an individual does not experience an event during the interval is the survival function $e^{-\lambda t}$, which can be derived from the Poisson distribution. The probability of exactly one event is

$$P(N(t)=1)=\frac{(\lambda t)^1 e^{-\lambda t}}{1!}=\lambda t e^{-\lambda t}\,. \tag{8}$$

If the times between two consecutive events are independent and exponentially distributed, then the waiting time to the *n*-th event, W_n, is the sum of n exponentially distributed arrival times:

$$W_n=\sum_{i=1}^{n} T_i\,, \tag{9}$$

which has a gamma distribution (Çinlar 1975, pp. 81-83). For the integer values of *n* it has the following probability density function

$$f(w_n)=\frac{\lambda^n e^{-\lambda w_n} w_n^{n-1}}{(n-1)!}\,,\ w_n>0\,. \tag{10}$$

The basic indicator of duration is the expected waiting time to migration, which can be also interpreted as an expected duration between successive migrations. It is

$$E[T_i] = \int_0^\infty S(t)dt = \int_0^\infty e^{-\lambda t} dt = \frac{1}{\lambda}. \tag{11}$$

Using the fact that the expectation of a sum is the sum of the expectations we obtain the expected waiting time till *n*-th migration

$$E[T_n] = \frac{n}{\lambda}. \tag{12}$$

The median waiting time (*me*) to migration, the time at which there is a probability of 50% that the migration took place, can be derived from the relationship

$$F(me) = 1 - e^{-\lambda me} = 0.5. \tag{13}$$

The median waiting time is therefore

$$me = \frac{\ln 2}{\lambda}. \tag{14}$$

For the exponential model of duration the median is smaller than the mean. It amounts to about 69.3% of the mean.

The basic Poisson process may be generalized by allowing λ to vary in time and to differ between subpopulations. These extensions are extensively covered in the literature. Although they are beyond the scope of this paper, some are presented briefly below in order to show possible areas of further research. The assumption that that each person faces the same constant hazard rate λ is restrictive. To take the differences between individuals into account we can introduce covariates in the model. If we model heterogeneity due to observable covariates, then the multiplicative hazards model due to Cox (1972), often called a proportional hazards model, is the most widely used one. A baseline hazard $\lambda_0(t)$ is multiplied by the term dependent only on the covariates. We first assume that the baseline hazard is time-constant and equal to λ. It means that each individual has constant intensity during lifetime, provided his or her characteristics are time-invariant. Thus we hold the underlying assumption on exponentially distributed interarrival times and Poisson distribution for counts. We receive the following model:

$$\lambda(t|x) = \lambda \cdot \exp(x'\beta), \tag{15}$$

where x is a column covariate vector and β is a corresponding coefficient vector.

Nonetheless, not all attributes of individuals are measured. Their observed characteristics may not be sufficient to completely capture the difference in intensity between them. Usually there exists an additional unobserved heterogeneity. It may be represented by a random variable Z whose realization z influences the hazard rate in a

multiplicative way. Given heterogeneity with respect to both observed and unobserved characteristics an intensity is then of the form

$$\lambda(t|x,z)=\lambda\cdot z\cdot\exp(x'\beta). \tag{16}$$

Since intensity is not negative, z must be limited to positive values, which restricts the potential distributions of Z. A random variable Z can follow a discrete or continuous distribution. If the distribution is discrete, the model is generally referred to as a mixture model. If the distribution is continuous, it is common to assume that $E(Z)=1$, i.e. the average individual faces the baseline hazard λ. The gamma distribution has a prominent role in modelling a positive continuous random effect. In some application random variable Z is supposed to capture the whole variability of hazard rate. The reason of omitting the covariates is the limitation of available data.

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 in time. Distributions that capture duration dependence of the event occurrence include among others 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. If the transition rates differ among heterogeneous individuals and the time dependence of the rate does not, the following proportional hazard model results:

$$\lambda(t|x,z)=\lambda_0(t)\cdot z\cdot\exp(x'\beta) \tag{17}$$

with time-dependent baseline hazard rate $\lambda_0(t)$.

The specification of the count model that is consistent with an assumed interarrival time distribution is not straightforward. In this paper we limit ourselves to a general relation. Let T_n denote the arrival time of the *n*-th event and let $N(t)$ represent the total number of events in the interval $(0,t)$. Then, the relationship between waiting times and the number of events is

$$N(t)<n \Leftrightarrow T_n \geq t\,. \tag{18}$$

Thus,

$$P(N(t)=n)=P(N(t)<n+1)-P(N(t)<n)=P(T_{n+1}\geq t)-P(T_n\geq t)=F_n(t)-F_{n+1}(t), \tag{19}$$

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 cumulative distribution function of the sum of *n* waiting times. Unlike in the case of

exponential distribution, it is usually not easily solved for other interarrival time distributions (see Bradlow et al. 2006 for Weibull distribution, and Winkelmann 1995 for gamma distribution).

3. Observation plans, measurements, and statistics

The migration process is a continuous and recurrent phenomenon. To collect data generated by a continuous-time process different observation plans, i.e. different schemes for collecting systematic information, can be used (Blossfeld and Rohwer 2002, Tuma and Hannan 1984). The most detailed information one can get on the process under study is information about all migrations of individuals with their exact timing. If they are available for the whole individual lifetimes then they are the most complete data possible. Recall that the migration history of an individual can be viewed from two different, but closely related, perspectives. In the first, the migration history is described in terms if the events and their timing (event approach). In the second, the migration history is described in terms of the places of residence at consecutive ages (status approach). (Rajulton 2001) provides a direct connection between event and status approach, defining an event as a transition between statuses (states). In the theory, when the continuous time scale is used, the two approaches are equivalent. Complete migration histories of all members of the population are, however, never available and in some instances event data are available while in other cases status data exist (with intervals of different lengths). In addition to incompleteness of collected data, the production of migration statistics involves aggregation of different data types. In result data available to end user are far from being complete. This section proposes a useful typology of existing migration data.

The distinction between *migration data* and *migrant data* is an established one in migration statistics. 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. It is clear that due to possible multiple migrations the total number of migrations is always greater than or equal to the total number of migrants. The number of migrants is often estimated through a census or survey question on the place of residence at a previous date, thus based on *status data*. As indicated by Courgeau (1979) this estimation is not satisfactory because return 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 event are sometimes defined as transition between statuses, for 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 following separate categories: *migration data*, *migrant data* (as defined by Courgeau 1973, 1979), and *discrete transition data* (hereinafter referred to as *transition data*). Usually migration data constitute the broadest category. Migrations are, however, not always observed and recorded in continuous time. Continuous registration of migration events is characteristic for population registers. Nonetheless, migration data can be also collected in census and surveys. Then usually occurrence of one migration in a reference period is recorded, most often it is the last migration.

We now introduce 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). 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: *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. In practice, as mentioned in the introduction, the duration may be intended or actual, but in this study we focus on the latter type. 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*. They cover persons who change their country of usual residence for a period of at least a year. It goes without saying that the longer the duration of stay threshold the less migrations and migrants are counted.

In sum, for the same underlying data generation process we receive different results depending on how the data happened to be collected and how the statistics happened to be produced. Discrepancies between different measures can be large. However, in order to understand the phenomenon and to compute diverse quantities of interest an adequate representation of the underlying process is needed. Thus, first we should identify what the underlying process is and then try to estimate its parameter based on available, usually incomplete, data.

4. Indicators of migration process

In this section we link empirical migration measures and migration intensities, which are the parameters of the migration process. Assume that members of a population migrate

independently and that their migration experience may be described by the same probability model. We further assume a simple probability model with migration intensities that do not vary in time and do not vary between individuals. The model and the associated migration indicators are given in Section 2. In this section, we start with movement approach and consider the *conditional migration measures* and *conditional migrant measures*. Next we consider the relation between the two. Finally, we present *transition data* and compare them with data produced based on movement approach.

Counting all migration movements, 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 refers to a reference period). In practice, however, only selected relocations are counted as migrations. Most often there are some minimum time constraints imposed on the length of stay that follows change of place of residence to distinguish migration from other types of relocations, which leads to *conditional migration data.* 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. If the migration rate is constant, then the probability of being a stayer at t_m is the survivor function of the exponential distribution or zero term in Poisson distribution. Therefore, an expected number of conditional migrations experienced by an individual over a period of specified length t is

$$E\left[N_{t_m}(t)\right]=\left[\sum_{n=0}^{\infty} n\,P\left(N(t)=n\right)\right]S(t_m)=\left[\sum_{n=0}^{\infty} n\cdot\frac{(\lambda t)^n e^{-\lambda t}}{n!}\right]e^{-\lambda t_m}=\lambda t e^{-\lambda t_m}, \quad (20)$$

where t_m is a time criterion used in a migration definition. The expected number of conditional migrations during the period from 0 to t is a proportion of the expected number of migrations. The expected number of migrations is given by the expression in brackets. The proportion of migrations that satisfy the duration of stay criterion is given by the survivor function $S(t_m)=e^{-\lambda t_m}$. Note that no restrictions have to be imposed on relation between lengths of reference period t and duration of stay t_m used to define migration. Thanks to stochastic approach we know the chances of staying for various durations t_m, even if the actual realisations take place beyond the reference period. In the special case $t_m=0$ equation (20) reduces to (7) for counts of all changes of places of residence. From (20) we obtain an important relation between counts of *conditional migrations* for different durations of stay, t_{m_1} and t_{m_2}:

$$E\left[N_{t_{m_1}}(t)\right]\Big/E\left[N_{t_{m_2}}(t)\right]=e^{-\lambda\left(t_{m_1}-t_{m_2}\right)}, \quad (21)$$

which is independent of the length of the reference period *t*. Different duration of stay criteria produce different migration figures and the discrepancy depends on the migration intensity. We estimate the difference for different migration intensities. Since one year duration is recommended by the United Nations and at the same time required by the EU Regulation, we use it as a reference level. We calculated the values of the ratio (21) for different durations applied in migration definition, $t_{m_1} \in [0;5]$, relative to the UN definition, $t_{m_2} = 1$, and for various migration intensity levels, $\lambda \in (0;1]$. The results for all migration intensities are presented in the left panel of Figure 1, but for clarity reasons the line graph in the right panel shows the values of the same ratio for selected λ. The colour scale (left panel) and the value axis (right panel) show the ratio of the number of conditional migrations based on the duration of stay criterion $t_m = t_{m_1}$ and the number of conditional migrations based on the duration of stay criterion used in the UN definition (one year), $t_{m_2} = 1$. For instance, if migration intensity equals 0.2 and we count migrations for half a year, $t_{m_1} = 0.5$, instead of one year we report figures that are higher by around ten percent. For the same migration rate of 0.2, counting migrations for two years, $t_{m_1} = 2$, results in an underestimation of the statistics by approximately 18%. In the left panel the area between the two contour lines includes combinations of migration intensities and duration of stay criteria for which the number of recorded migrations differs by no more than ten percent compared to the number of migrations involving a stay of minimum one year. Note that for the low levels of relocation intensities discrepancies between counts of migrations for different durations are relatively small. The higher the migration rate the larger the differences. If the migration intensity is high, the numbers of migrations recorded under a duration of stay criterion that differs from one year can be considerably different from the number recorded under the one year duration of stay criterion and the difference increases rapidly when the criterion diverges from the one year criterion. It results from the fact that with the increasing intensity a person relocates more often. In other words it means that 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.

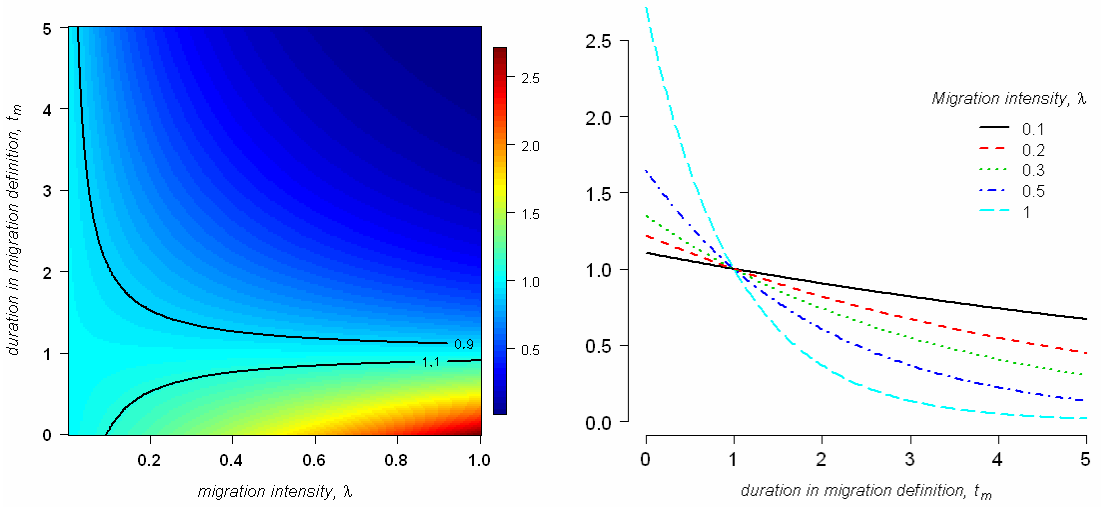

FIGURE 1. Ratio of migrations for various duration t_m to migrations for one year (right panel: selected intensities)

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 and count only migrants who experienced at least one migration followed by a stay of specified duration. Note that, as described in Section 2, the concept of *conditional migrant data* differs from the concept of *discrete transitions*. Consider a person who migrates two times during a reference period of one year. In the former case he or she is counted if one of the relocations is followed by a stay of the duration in question. In the latter case a person is included in the 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 can not be a return one. We calculated ratios analogous to (21) 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. They are shown in Figure 2, which is analogous to Figure 1 for *conditional migrations*. The left panel shows for different migration intensities, the ratio of the number of conditional migrants for the duration of stay criterion $t_M = t_{M_1}$ and the number of conditional migrants for the duration of stay criterion used in the UN definition (one year), $t_{M_2} = 1$. The right panel shows the same ratio for selected values of migration intensity.

Note that unlike for *conditional migration data* discrepancies between *conditional migrant data* 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 one year nor for five years may be experienced more than once within one year period. Hence, differences between migrant data for one year stay and five-year stay respectively are exactly the same as in the case of migration data. Within a three-year period multiple migrations are possible in the case of migration for one year but not for five years. In result, the multiple migrations that are not included in statistics on migrants diminish the discrepancy between one year and five years conditional migrant data compared to 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 number of multiple migrants is negligible (see Figure 3).

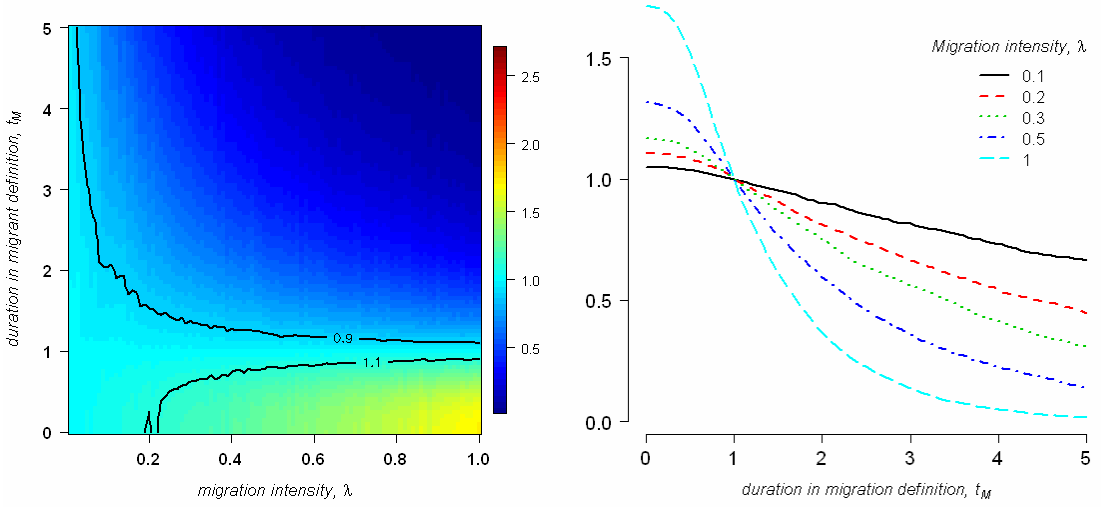

FIGURE 2. Ratio of migrants for various duration t_M to migrants for one year; annual data (right panel: selected intensities)

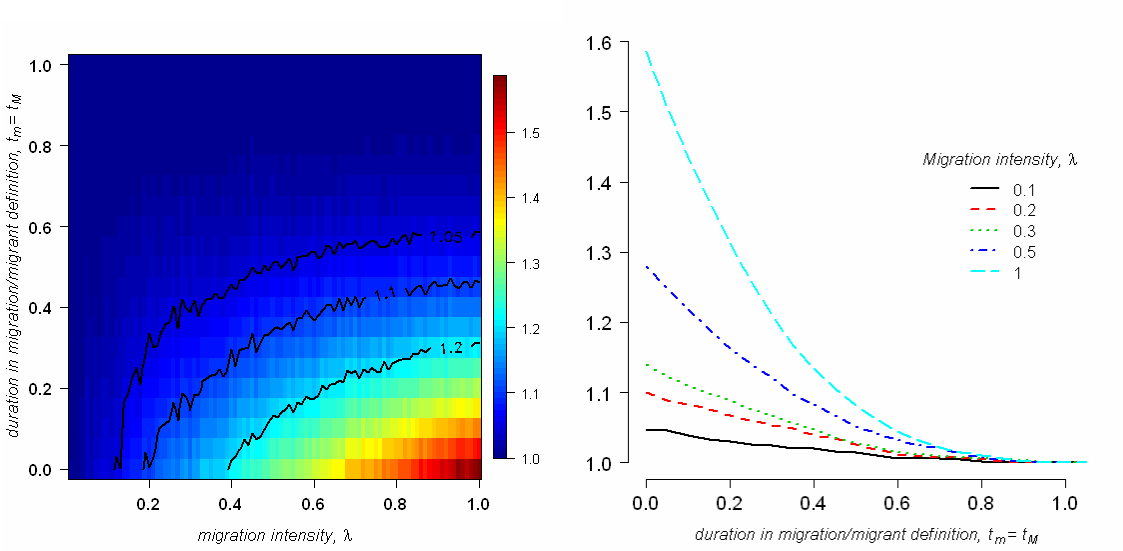

FIGURE 3. Migrations per migrant (right panel: selected intensities)

Our analysis shows that, under the simplified assumptions on migration process and provided we know the relocation rate, we can recalculate counts of migrations or migrants for a specific duration (conditional migration and migrant data) into migrations or migrants for any other required duration. An example of relations between these types of measures for durations up to one year and intensity $\lambda = 0.2$ is presented in Figure 4. The contour line of value one indicates durations of stay t_m and t_M used in migration and migrant definition respectively which lead to the same level of reported migration flows. For instance, besides the obvious case of migrations and migrants for one year, the number of migrants for two 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 know number of migrations for the same or different duration we have to multiply our figure by the value indicated by the colour scale. Within one year duration limit the discrepancy between the narrowest and the broadest measure, namely number of conditional migrants for one year, $t_M = 1$, and number of all (non-conditional) migrations, $t_m = 0$, respectively, equals 22% (upper left corner of Figure 4). It means that, during a period of one year, the number of migrations without any duration of stay restriction is 22% larger 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 for duration up to one year, which are usually used in practice, we should not expect differences largest than 50%. Nonetheless, if the widest measure is the conditional migrants for five years, which may approximate the measure of permanent migrants applied by e.g. some former state socialist countries, the difference increases to 172% for intensity $\lambda = 0.2$. For 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, e.g. it amounts to 2% for $\lambda = 0.8$, but such high international migration rate is vastly unrealistic nonetheless.

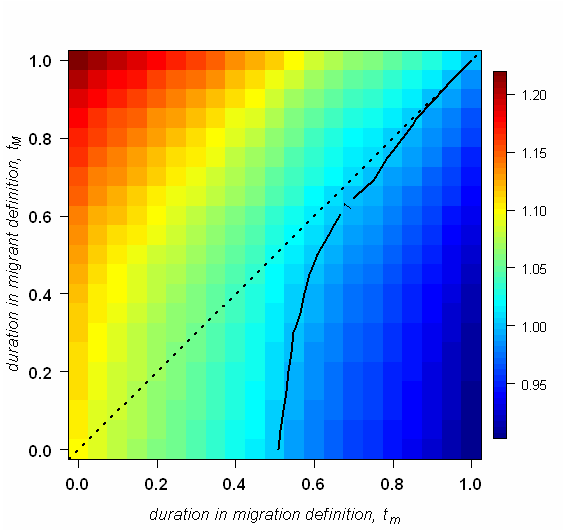

FIGURE 4. Ratio of migrations to migrants, for various durations up to one year and intensity λ=0.2

So far we considered *conditional migration* and *conditional migrant* measures, which are based on *movement approach*. These data types are predominant in European statistical practice. Most of the annual statistics on international migration flows produced in Europe represent one of them. Now we consider *transition approach*, i.e. *direct transition* measures that are based on the comparison of person's usual place of residence at two consecutive points in time. First, we compare numbers of transitions for intervals of different length. For instance we want to determine what proportion of transitions recorded over a five year interval would also be recorded if the interval is one year. Then, we compare transitions over intervals of different length and conditional migrations for various durations of stay.

Transition data are characteristic for census and household survey. 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 one year or five years prior to enumeration. Appropriate data are collected in many countries, 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 relation between different migration measures concentrate on this type of measures derived for time intervals of various lengths, e.g. one- and five-year period (see Kitsul and Philipov 1981, Liaw 1984, Long and Boertlein 1990, Rees 1977, Rogers et al. 2003, Rogerson 1990).

Consider, for illustration, a very simplified case when individuals migrate between two areas that form a closed system, with equal and constant intensity and migrations 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):

$$p = \lambda t e^{-\lambda t} + \frac{(\lambda t)^3 e^{-\lambda t}}{3!} + \ldots = \frac{1 - e^{-2\lambda t}}{2}. \qquad (22)$$

When a person migrates an even number of times he or she is in the same area at the beginning and end of the migration interval. These return moves as well as other repeat moves do not increase the number of transitions. Hence, even with the constant migration intensity, the total number of transitions does not increase linearly with time as it is the case for relocations. Figure 5 provides a graphical presentation of these changes for selected intensities (compare Rogerson 1990).

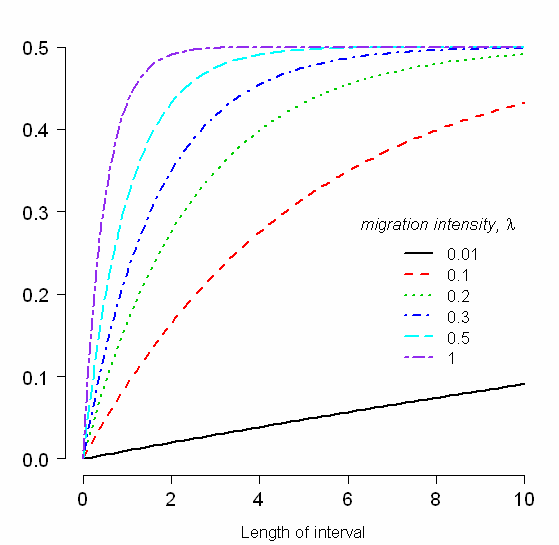

FIGURE 5. Discrete time transitions over intervals of different lengths for selected intensities (expected values per individual).

Thus, the number of transitions over *n*-year period is not equal to *n* times the number of transitions over one year. Based on (22) we have the following relation between numbers of transitions N_p over time intervals of different lengths denoted by t_{p_1} and t_{p_2}

$$E\left[N_p\left(t_{p_1}\right)\right] / E\left[N_p\left(t_{p_2}\right)\right] = \frac{1 - e^{-2\lambda t_{p_1}}}{1 - e^{-2\lambda t_{p_2}}}. \qquad (23)$$

Figure 6 shows ratio of transitions over a few-year periods to transitions over one year, depending on the level of migration rate. In general, the higher the intensity, the lower the discrepancies between measures. It results from the fact that increase in rate raises the chance of primary migration in short periods of time and repeat migrations in longer ones. Nonetheless, the extreme values of rates for which different measures are hardly distinguishable are presumably only theoretical. Consider transitions over a five year interval compared to transitions over one year. Empirical five-year to one-year ratios reported in the literature for the internal migration take on values between two and four (Long and Boertlein 1990, Rees 1977, Rogers et al. 2003). Since internal migration is more prevalent than international we can expect that values greater than four are quite realistic for international migration. They correspond with intensity λ lower than 0.06. At the same time ratios of values lower than two, corresponding to migration rate higher than 0.33, can be questionable.

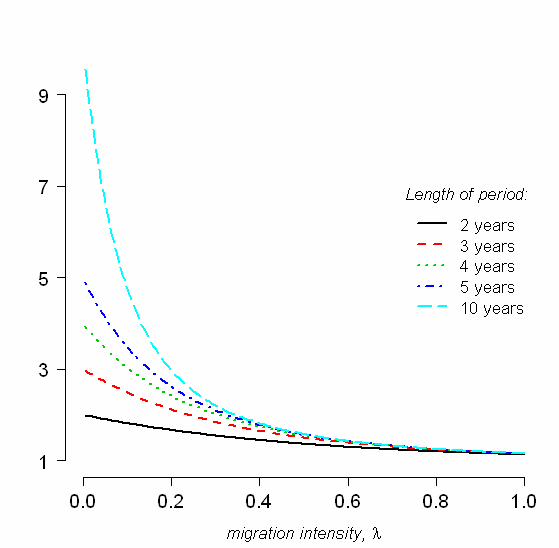

FIGURE 6. Ratio of transitions over an interval of different lengths to transitions over one year.

Under the simplified assumptions stated above it is equally easy to derive relation 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 reference period for migrations number is equal to interval over which we count number of transitions. The duration of stay criterion t_m used in migration definition may vary. For example, we compare number of migrations that take place during a reference year, $t = 1$, and are followed by at least half-year stay, $t_m = 0.5$, with number of people whose places of residence at the beginning and of the end of this reference year, $t = 1$, differ. From (20) and (22) we obtain

$$E\left[N_p(t)\right] / E\left[N_{t_m}(t)\right] = \frac{e^{\lambda t_m}\left(1 - e^{-2\lambda t}\right)}{2\lambda t}, \qquad (24)$$

which enable 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. Assuming different migration rates let us look at the 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 (see Figure 7). For low, and at the same time most realistic, migration intensities the differences between these measures are negligible. For higher levels of rate transition approach leads to larger migration measure. In other words, for one-year interval the number of transitions is higher than the number of conditional migrations that are followed by one year stay.

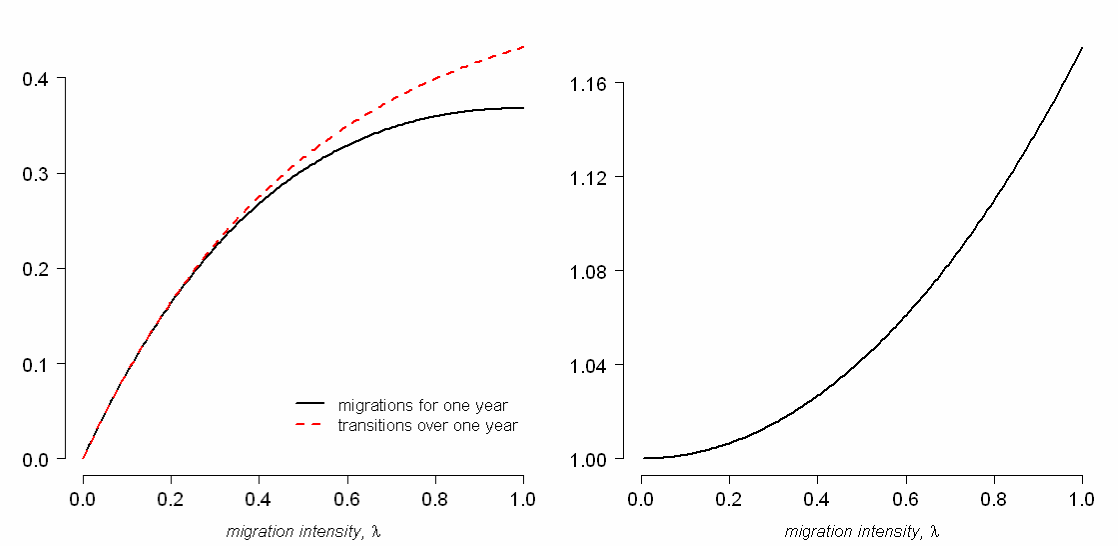

FIGURE 7. Migrations for one year and transitions over one year: expected values per individual (left) and ratio between them (right).

In general, the transition approach includes less information than the movement approach, because observation of the places of residence of individuals at the beginning and end of an interval ignores multiple and return migrations within a time interval. In the case of annual measure on conditional migration for one year, however, multiple and return migrations are not possible and number of migrations is equal to number of migrants. Moreover, in the transition approach no duration criterion is imposed on the stay following relocation and this explains the larger measure of migration that results.

5. Conclusions

We have illustrated how theory of stochastic processes may yield important insights into the problems of inconsistency of migration statistics. The main focus has been put on the time criterion used in migration measure to select migrations from all changes of place of residence. The time refers to duration of stay following relocation, which is specified very differently among countries and constitutes the main source of discrepancies in operationalization of migration concept in the EU countries. The application of the simple homogenous Poisson model shows that it inevitably leads to different numbers of recorded migration. The level of discrepancies depends on the migration rate. Under the simplified assumptions in the model a straightforward relation exists between migration measures used in common migration statistics and the migration intensity, which is defined unambiguously. The Poisson model used in this study for illustration purposes does not give an accurate description of migration. It may be treated as a point of departure for further research in this area. The research should aim at applying a generalized model of a counting process that accounts for non-constant transition intensities and population heterogeneity. Section 2 of the paper introduced the subject.

References

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. (1993) *Statistical models based on counting processes,* New York, Springer-Verlag.

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.

Bilsborrow, R. E., Hugo, G., Oberai, A. S. & Zlotnik, H. (1997) *International migration statistics: guidelines for improving data collection systems,* Geneva, International Labour Office.

Blossfeld, H. P. & Rohwer, G. (2002) *Techniques of event history modeling: new approaches to causal analysis,* New Jersey, Lawrence Erlbaum Associates.

Bradlow, E., Fader, P., Adrian, M. & Mcshane, B. (2006) *Count Models Based on Weibull Interarrival Times*, SSRN.

Çinlar, E. (1975) *Introduction to stochastic processes*, Englewood Cliffs, New Jersey: Prentice Hall, Inc.

Courgeau, D. (1974) Methodological Aspects of the Measurement of International Migration. In G. Tapinos (Ed.) *International Migration Review.* Paris, Committee for International Coordination of National Research in Demography.

Courgeau, D. (1979) Migrants and migrations. *Population, Selected Papers,* 3, 1-35. (French version published in 1973).

Cox, D. R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological),* 34, 187-220.

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.

Ginsberg, R. B. (1971) Semi-Markov processes and mobility. *Journal of Mathematical Sociology,* 1, 233-262.

Ginsberg, R. B. (1972) Critique of probabilistic models : Application of the semi-Markov model to migration. *Journal of Mathematical Sociology,* 2, 63-82.

Ginsberg, R. B. (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, R. B. (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 A. Rogers (Ed.) *Advances in multiregional demography* Research Report RR-81-6, Laxenburg, Austria, International Institute for Applied Systems Analysis.

Klein, J. P. & Moeschberger, M. L. (2003) *Survival analysis: techniques for censored and truncated data*, Springer.

Kupiszewska, D. & Nowok, B. (2008) Comparability of statistics on international migration flows in the European Union. In J. Raymer & F. Willekens (Eds.)

International Migration in Europe: Data, Models and Estimates. Chichester, John Wiley & Sons, Ltd.
Lancaster, T. (1990) *The econometric analysis of transition data,* New York, Cambridge University Press.
Ledent, J. (1980) Multistate life tables: movement versus transition perspectives. *Environment and Planning A,* 12, 533–562.
Liaw, K.-L. (1984) Interpolation of transition matrices by the variable power method. *Environment and Planning A* 16, 917-925.
Long, J. F. & Boertlein, C. G. (1990) Comparing migration measures having different intervals. *Current Population Reports* Washington, U.S. Census Bureau.
Nowok, B., Kupiszewska, D. & Poulain, M. (2006) Statistics on international migration flows. In M. Poulain, N. Perrin & A. Singleton (Eds.) *THESIM: Towards Harmonised European Statistics on International Migration.* Louvain-la-Neuve, Presses Universitaires de Louvain.
Poulain, M. (1999) International migration within Europe: Towards more complete and reliable data. *Working Paper No. 37. Joint ECE-Eurostat Work Session on Demographic Projections.* Perugia.
Poulain, M. (2001) Is the measurement of international migration flows improving in Europe? *Working Paper No. 12, Joint ECE-Eurostat Work Session on Migration Statistics.* Geneva.
Poulain, M., Perrin, N. & Singleton, A. (Eds.) (2006) *THESIM: Towards Harmonised European Statistics on International Migration,* Louvain-la-Neuve, Presses Universitaires de Louvain.
Rajulton, F. (2001) Analysis of life histories: a state space approach. *Canadian Studies in Population,* 28, 341-359.
Rees, P. & Willekens, F. (1986) Data and accounts. In A. Rogers & F. Willekens (Eds.) *Migration and settlement: a multiregional comparative study.* Dordrecht, Reidel Press
Rees, P. H. (1977) The measurement of migration, from census data and other sources. *Environment and Planning A,* 9, 247-272.
Rogers, A., Raymer, J. & Newbold, K. B. (2003) Reconciling and translating migration data collected over time intervals of differing widths. *Annals of Regional Science,* 37, 581-601.
Rogerson, P. A. (1990) Migration analysis using data with time intervals of differing widths. *Papers in Regional Science,* 68, 97-106.
Tuma, N. B. & Hannan, M. T. (1984) *Social dynamics: models and methods*, Academic Press.
United Nations (1998) *Recommendations on Statistics of International Migration: Revision 1,* New York, United Nations (Statistical Papers, No. 58, Rev.1 Sales No. E.98.XVII.14).
United Nations (2002) *International Migration Report 2002,* New York, United Nations Population Division, Department of Economic and Social Affairs.
Willekens, F. (1982) Identification and measurement of spatial population movements. *ESCAP, National Migration Survey Series Manual No. X: Guidelines for Analysis.* New York, United Nations.

Willekens, F. (1985) Comparability of migration data: Utopia or reality? . In M. Poulain (Ed.) *Migrations internes. Collecte des données et méthodes d'analyse.* Cabay, Louvain-la-Neuve.

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.