

Models of Migration by Age and Spatial Structures

Frans Willekens

Netherlands Interdisciplinary Demographic Institute (NIDI)
The Netherlands

Paper prepared for presentation at the **Colorado Conference on the Estimation of Migration**, 24 – 26 September 2004

1. Introduction

The paper provides a unified perspective on the modeling of migration. In applied research, the level of migration is measured in different ways: number of migrants during a period of time, the share of migrants in a population, and the rate of migration. The different measures are related in some way. A unified perspective should encompass different measures of migration levels in a coherent framework. In migration studies, a frequently made distinction is between migration and migrants (Courgeau, 1973). Migration is the *event* of relocation (beyond administrative boundaries). Migrant refers to a *person* who relocates. In migrant data, relocation is measured by comparing the residence at two points in time, a fixed or variable number of years apart (one year, five years, lifetime). The unified perspective should encompass a way to convert migration data into migrant data and vice versa. Differences in interval lengths have occupied researchers for years (see e.g. Long and Boertlein, 1981; Kitsul and Philipov, 1981, Courgeau, 1982).

The unified perspective should also distinguish characteristics of migrations (events) and characteristics of migrants (persons). Characteristics of the event include the origin and destination of migration, and the reason for migration. Examples include rural-urban migration, international migration by country of origin and country of destination, marriage migration, and job-related migration. Note that the origin is the current place of residence, which is also a characteristic of the person. Age, sex, level of education, marital status, employment status, country of birth and country of residence at a given age are characteristics of migrants. In the present paper, and in life history analysis in general, age is treated differently from the other characteristics. Age is a duration variable. In general, a duration variable is a variable that measures the time elapsed since a reference event or event-origin. Any life event can be selected as the event-origin (e.g. birth, marriage, last migration).

The unified perspective is embedded in a life course perspective on migration. At a given age, an individual has a multitude of **attributes**, characteristics or traits: gender, marital status, maternal status [presence/absence of children or number of children], migrant status (ever/never migrated), place of residence, living arrangement, health status, employment status, educational status, source of income, etc. The attributes relate to domains of life such as marriage, parenting, and employment. An individual may be viewed as a carrier of attributes. Most attributes considered in demographic studies have a finite number of categories. Hence, attributes are represented by discrete or categorical variables.

Personal attributes vary with age and in time. A change in attribute is an **event**. The history or sequence of attributes during a given period or a lifetime is fully documented if the attribute of the individual is known at every age. An alternative but equivalent approach is to record the initial attribute, the age or time when an attribute changes (i.e. time at event) and the new attribute after the change (direction of change). The first is the *status-based approach*, the second the *event-based approach* to life history recording.

The status-based approach is central to multistate analysis of life histories; the event-based approach is central to event history analysis. Although attributes may change at any age or any point in time, the measurement of the change may be restricted to time intervals. Events are said to occur in continuous time whereas the occurrence may be recorded in continuous time (exact time) or discrete time (time in completed months or years; time intervals). In the proposed framework, the distinction between continuous time and discrete time is essential. In the literature, the measurement of events in continuous time has been referred to as the movement approach and the measurement in discrete time as the transition approach (Ledent, 1980).

In multistate life history models, the attribute variable is known as the state variable and a particular value of the categorical variable as the state occupied. The possible combination of states or categories is state space. In this paper, we adopt a multistate modeling perspective. The multistate modeling framework is rooted in the multivariate analysis of time-to-event data, also known as multistate survival analysis. In multistate analyses two key concepts are distinguished. The first is *state occupancy*. It is the attribute or the state occupied at a given point in time (e.g. at a given exact age). The second is the *state transition*. It refers to a change in attribute or state occupied. Transitions may be expressed in continuous time and discrete time (Andersen et al., 1993, p. 93). Transitions in continuous time are referred to as *direct transitions* or events (Rajulton, 1999, p. 5). *Discrete-time transitions* are measured by comparing states occupied at two consecutive points in time. A person who resides in a given region at t_1 and in a different region at t_2 has made at least one migration (experienced at least one event).

Migration may be measured in various ways. Hence, several data types may be distinguished. The data may relate to events (continuous time) or to changes in status (discrete time). They may be individual data (micro-data) or aggregate data (tabulated data, contingency tables). Aggregated data are grouped data. Grouped data pertain to events or persons that are grouped on the basis of particular characteristics. For instance, micro data from a sample of individuals may indicate the state each respondent occupies. Grouped data indicate the number of respondents in each of the states. In addition, the level of migration may be expressed in different ways. They may be grouped into three broad types, however: counts, probabilities and rates.

The characteristics of migrations and migrants are specified at the individual level. At the population level, the distribution of characteristics result in *data structures*: the age structure, the covariate structure, the motivational structure, and the spatial structure. The covariate structure relates to the characteristics of the migrant, e.g. country or region of birth and sex, but also employment status and marital status. The motivational structure relates to the reasons for migration. The spatial structure relates to the direction of migration. Each structure calls for a different modeling approach that is however logically integrated in the comprehensive framework. The age structure is modeled by models of age (duration) dependence. The model migration schedules are examples. The covariate structure is described by transition rate models and logistic regression models. The spatial structure is captured by spatial interaction models. In the unlikely case that

the four structures are independent, the structures can be modeled separately. Different dependencies or interaction effects are identified and integrated in the framework. For instance, the age structure of migration may differ by destination and/or covariate and/or reason for migration.

A unified perspective has three significant advantages. First, it provides a single, comprehensive framework for the *analysis of data* on migration. Second, it provides a framework for the *harmonization of migration statistics*. Migration data come in many different ways. The estimation of comparable indicators of levels and trends of migration (and direction of migration) and the comparative analysis of patterns of migration require comparable data or techniques for converting data types. The framework encompasses techniques for converting migrant data recorded for different time periods into migrant data pertaining to the same time period (one year, say). It also converts migrant data into migration data and enables the estimation of number of migrants from data on migrations, irrespective of the length of the period considered.

Third, it provides a unified framework for the *prediction (estimation) of missing data on migration*. The prediction of a missing value or a set of missing values on the basis of available data is similar to the imputation of missing values. Imputation is receiving much attention in the literature. Methods for statistical data imputation may be divided into two broad groups: model imputation and donor imputation. In model imputation, the imputed values are directly derived from a data model, i.e. a statistical or demographic model of the data. Common data models take the format of regression models. The regression model is estimated from the available data that may be augmented with qualitative information on migration (expert opinion, judgmental data). In donor imputation, the imputed values are derived from a set of observed values from real respondents (donors). The imputed value is based on information in the closest valid record (nearest neighbour matching characteristics that are not missing). Post-imputation edits (post-editing) make sure the nearest neighbour is close enough to be used as a donor. All available auxiliary information is used to assure a best estimate or imputation. The estimation of missing migration data may benefit from the literature on imputation. Donor imputation is not considered in this paper.

The unified framework for the modeling of migration is still incomplete. Three limitations are singled out. First, the method does not offer a measure of the reliability of the estimates or the degree of confidence one may attach to the estimates. Second, it does not cover the indirect estimation of migration from data on populations at two points in time and natural increase during the period. Third, it is of no use to estimate undocumented migration if undocumented migrants are not included in the aggregate data or qualitative information (e.g. expert opinion, educated guess) is lacking.

The paper consists of two main sections. Section 2 presents several probability models of migration emphasizing the interrelation. Different models are associated with different data types. Harmonization of migration data and comparison of migration levels in space and time require that one data type can be converted into another type. Probability models serve that function. Section 3 briefly presents a general method for model

estimation when data are incomplete. The method, which is widely used in migration analysis, is presented as a special case of the EM algorithm. The EM algorithm is the most widespread statistical method for model estimation when data are incomplete. The paper demonstrates that the modeling of migration can benefit from recent developments in modeling of life events and life histories.

2. Probability models of migration

This section reviews models of state occupancies and state transitions. State occupancy is expressed in terms of the probability that an individual selected at random from a (sample) population occupies a given state. It is approximated by the proportion of the (sample) population in a given state. State transitions during an interval (in continuous time or discrete time) are expressed in terms of three risk indicators: counts, probabilities (proportions) and rates. Probabilities relate the transitions to the *risk set* (population at risk) at the beginning of an interval. The risk set accounts for attrition during the interval for reasons unrelated to the transition being studied (censoring), i.e. the migration from the current place of residence to a given destination. Rates relate the transitions to the *exposure* to the risk of transition. Probabilities are obtained as the ratio of the number of events (in continuous time or discrete time) to the risk set; rates are obtained as the ratio of the number of events (in continuous time) to the exposure time. Counts refer to the number of events that occur during a given interval. It is the numerator in probability and rate measures.

In developing probability models of migration and migrants, the occurrence of an event (migration) is assumed to be the result of an underlying random mechanism. The occurrence of a migration depends on both personal attributes (systematic factors) and chance. Our approach is to model the random mechanism by specifying a probability model. A problem is that the event is not associated with a single random mechanism. Different mechanisms may result in the same event of migration. Hence different probability models may describe the level and direction of migration. It should be possible to identify the set of plausible mechanisms and to identify the mechanism that *most likely* produced the event. To determine the most likely mechanisms and the model that describes that mechanism, the maximum likelihood method is applied. The method identifies the ‘best model’, i.e. the model that has the greatest probability of predicting the observations on event occurrence. That model describes the random mechanism that most likely underlies the event.

2.1 State probabilities

Let \mathbf{S} denote the state space: $\mathbf{S} = \{1, 2, 3, \dots, i, \dots, I\}$. The state space contains I possible places of residence. At a given age, an individual resides in one place and one place only. In other words, the states are mutually exclusive. Let $Y_k(x)$ be a polytomous random variable denoting the state occupied by individual k at exact age x . The probability that an individual resides in state i is the *state probability*. The probability that individual k resides in state i at exact age x is ${}_k\pi_i(x) = \Pr\{Y_k(x)=i\}$. If all individuals are independent and identical, ${}_k\pi_i(x) = \pi_i(x)$ for all k . If individuals differ in a few characteristics only or if

a few characteristics suffice to predict the state occupied at age x , then ${}_k\pi_i(x) = \pi_i(x, Z)$, where Z represents a specific combination of characteristics or covariates. The probability that individual k occupies state i at exact age x depends on the covariates only and individuals with the same covariates have the same state probability.

The state occupied at x may be denoted differently. Let $Y_{ki}(x)$ be an indicator variable (binary) which is 1 if individual k occupies state i at x and 0 otherwise. The state probability is the probability that the random variable takes on the value of 1.

Consider a sample of m individuals. We do not consider covariates, implying that all individuals are identical. Covariates are introduced below. In addition, age is omitted for convenience. The number of individuals observed in state i is

$$N_i = \sum_{k=1}^m Y_{ki}$$

The probability of observing N_1 individuals in state 1, N_2 in state 2, N_3 in state 3, etc., is given by the multinomial distribution

$$\Pr\{N_1 = n_1, N_2 = n_2, \dots\} = \frac{m!}{\prod_{i=1}^I n_i!} \prod_{i=1}^I \pi_i^{n_i}$$

where n_i is the observed number of individuals in i and with $\sum \pi_i = 1$ and

$\sum N_i = \sum n_i = m$. The most likely values of the parameters π_i , given the data, are obtained by maximizing the likelihood that the model predicts the data, which is the maximum likelihood method. The values of π_i ($i = 1, 2, \dots, I$) that maximize the above multinomial distribution is $\hat{\pi}_i = \frac{n_i}{m}$

The expected number of individuals occupying state i is $E[N_i] = \pi_i m$ and the variance is $\text{Var}[N_i] = \pi_i(1-\pi_i)m$. The probability that an individual is found in state i is the expected value of Y_i : $\pi_i = E[Y_i]$. The variance of Y_i is $\text{Var}[Y_i] = \text{Var}[N_i/m] = \text{Var}[N_i]/m^2 = [\pi_i(1-\pi_i)]/m$. The variance declines with increasing sample size.

Now we introduce covariates. They are denoted by Z ($Z = \{Z_1, Z_2, Z_3, \dots\}$). Z_p may represent a single attribute or a combination of attributes (to denote interaction effects). The state probability $\pi_i(Z)$ that an individual with covariates Z occupies state i is given by the logit equation

$$\log \text{it}(\pi_i) = \ln \frac{\pi_i}{1-\pi_i} = \eta_i = \beta_{i0} + \beta_{i1}Z_1 + \beta_{i2}Z_2 + \beta_{i3}Z_3 + \dots$$

where $\frac{\pi_i}{1-\pi_i}$ is the odds of occupying state i . The logit transformation assures that the state probabilities lie between 0 and 1, and that their sum is equal to one. The value of η

may range from $-\infty$ to $+\infty$, but the value of π_i stays within 0 and 1. To obtain the probabilities, the logit scale is converted into the probability scale:

$$\pi_i = \frac{\exp(\eta_i)}{\exp(\eta_1) + \exp(\eta_2) + \dots + 1 + \dots} = \frac{\exp(\eta_i)}{\sum_{j=1}^I \exp(\eta_j)}$$

where the 1 is associated with the reference category. The model is the multinomial logistic regression model.

2.2 Transition probabilities

The state occupied at a given age generally depends on the states occupied at previous ages, in addition to personal attributes at the given age. Hence the probability of being in state j at $x+1$ (or more generally y) depends on the states occupied at previous ages. It is often assumed that only the most recent state occupancy is relevant:

$$Pr\{Y(x+1) = j / Y(x), Y(x-1), \dots; Z\} = Pr\{Y(x+1) = j / Y(x); Z\}$$

If the state occupied at x is i , then

$$Pr\{Y(x+1) = j / Y(x) = i\} = p_{ij}(x)$$

$p_{ij}(x)$ is the probability that an individual who resides in state i at x resides in state j at $x+1$. It is the discrete-time transition probability. The interval can be of any length but is generally one or five years. This model is suited for describing migrant data, i.e. data that infer migration by recording the places of residence at two consecutive points in time. The method is related to the Option 2 method (Rogers, 1975).

The status dependence may also be written as

$$\log it[\pi_j(x+1)] = \beta_{j0} + \beta_{j1} Y_i(x)$$

where $Y_i = 1$ if state i is occupied at x and 0 otherwise. Hence the transition probability may be written as

$$p_{ij}(x) = \frac{\exp[\beta_{j0} + \beta_{j1} Y_i(x)]}{\sum_{r=1}^I \exp[\beta_{j0} + \beta_{j1} Y_r(x)]}$$

The transition probabilities may depend on covariates in a way that is similar to that of state probabilities. Transition probabilities out of a given state i that depend on covariates may be estimated using multinomial logistic regression software.

2.3 Transition rates

The transition probabilities discussed in Section 2.2. are defined for discrete time intervals. They depend on the number of persons at risk at the beginning of the interval, which is generally known as the risk set. The probabilities are not directly related to the

duration that individuals in i are at risk of migrating to j . Since the event of migration (direct transition) may occur at any time during the interval from x to y (with $y = x+1$, for instance), the transition probability is defined for very small intervals. The probability that an individual in i transfers to j during an infinitesimally small interval following x is the instantaneous rate of transition:

$$\mu_{ij}(x) = \lim_{(y-x) \rightarrow 0} \frac{p_{ij}(x, y)}{y - x} \quad \text{for } i \text{ not equal to } j.$$

The instantaneous rate of transition is also known as the transition intensity and the force of transition. The term $\mu_{ii}(x)$ is defined such that

$$\sum_j \mu_{ij}(x) = 0$$

Hence

$$\mu_{ii}(x) = \sum_{j \neq i} \mu_{ij}(x) = \lim_{(y-x) \rightarrow 0} \frac{1 - p_{ij}(x)}{y - x}$$

The quantity $\mu_{ii}(x)$ is non-negative. It is sometimes referred to as the intensity of passage because it relates to the transition from i to any other state different from i . Schoen (1988, p. 65) refers to it as the ‘force of retention’.

The intensities are the basic parameters of a continuous-time multistate process. Under the restrictive Markov assumption, the probability that an individual leaves a state depends only on the state and the individual’s age. It is independent of other characteristics.

The matrix of instantaneous rates with off-diagonal elements $-\mu_{ij}(x)$ and with $\mu_{ii}(x)$ on the diagonal is known as the generator of the stochastic process $\{Y_k(x); x \geq 0\}$ (Çinlar, 1975, p. 256). The matrix is denoted by $\boldsymbol{\mu}(x)$. It has the following configuration:

$$\boldsymbol{\mu}(x) = \begin{bmatrix} \mu_{11}(x) & -\mu_{21}(x) & \cdot & \cdot & -\mu_{11}(x) \\ -\mu_{12}(x) & \mu_{22}(x) & \cdot & \cdot & -\mu_{12}(x) \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ -\mu_{11}(x) & -\mu_{21}(x) & \cdot & \cdot & \mu_{11}(x) \end{bmatrix}$$

Note that

$$\lim_{(y-x) \rightarrow 0} \frac{\mathbf{P}(x, y) - \mathbf{I}}{y - x} = -\boldsymbol{\mu}(x)$$

The matrix of discrete-time transition probabilities is:

$$\mathbf{P}(x,y) = \begin{bmatrix} p_{11}(x,y) & p_{21}(x,y) & \dots & p_{N1}(x,y) \\ p_{12}(x,y) & p_{22}(x,y) & \dots & p_{N2}(x,y) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ p_{1N}(x,y) & p_{2N}(x,y) & \dots & p_{NN}(x,y) \end{bmatrix}$$

An element of $\mathbf{P}(x,y)$, $p_{ij}(x,y)$, denotes the probability that an individual who is in state i at exact age x is in state j at exact age y . The Markovian assumption implies the following relationship between $\mathbf{P}(x,x+v)$ and $\mathbf{P}(x+v,y)$:

$$\mathbf{P}(x,y) = \mathbf{P}(x,x+v) * \mathbf{P}(x+v,y).$$

Subtraction of $\mathbf{P}(x+v,y)$ from both sides of the equation yields

$$\frac{\mathbf{P}(x,y) - \mathbf{P}(x+v,y)}{v} = \frac{[\mathbf{P}(x,x+v) - \mathbf{I}]\mathbf{P}(x+v,y)}{v}$$

and

$$\lim_{v \rightarrow 0} \frac{P(x,y) - P(x+v,y)}{v} = \lim_{v \rightarrow 0} [P(x,x+v) - I]P(x+v,y)$$

or

$$\frac{d\mathbf{P}(x)}{dx} = -\boldsymbol{\mu}(x)\mathbf{P}(x)$$

The model is a system of differential equations. To solve the system, the transition intensities are assumed to remain constant during the age interval from x to y and that the transition intensities during that age interval are equal to the empirical occurrence-exposure rates for that age interval. The matrix of transition probability from x to y is

$$\mathbf{P}(x,y) = \exp[-(y-x)\mathbf{M}(x,y)]$$

where $\mathbf{M}(x,y)$ is the matrix of empirical occurrence-exposure rates or transition rates for the age interval from x to y and $\mu_{ij}(t) = m_{ij}(x,y)$ for $x \leq t < y$ and $\boldsymbol{\mu}(t) = \mathbf{M}(x,y)$ for $x \leq t < y$.

A number of methods exists to determine the value of $\exp[-\mathbf{M}]$ (see e.g. Director and Rohrer, 1972, pp. 431ff; Aoki, 1976, p. 387; Strang, 1980, p. 206). We use the Taylor series expansion. Note that for matrix \mathbf{A} , we may write

$$\exp(\mathbf{A}) = \mathbf{I} + \mathbf{A} + \frac{1}{2!}\mathbf{A}^2 + \frac{1}{3!}\mathbf{A}^3 + \dots$$

Hence

$\exp[-(y-x)\mathbf{M}(x,y)] = \mathbf{I} - (y-x)\mathbf{M}(x,y) + \frac{(y-x)^2}{2!} [\mathbf{M}(x,y)]^2 - \frac{(y-x)^3}{3!} [\mathbf{M}(x,y)]^3 + \dots$ (see also Schoen, 1988, p. 72).

The transition probabilities may be approximated by assuming that the events are uniformly distributed during the interval from x to y . The discrete-time transition probability matrix under the linear model is given by the following expression:

$$\mathbf{P}(x,y) = \left[\mathbf{I} + \frac{1}{2} \mathbf{M}(x,y) \right]^{-1} \left[\mathbf{I} - \frac{1}{2} \mathbf{M}(x,y) \right]$$

The approximation is adequate when the transition rates are small or the interval is short.

The instantaneous rates of transition $\mu_{ij}(x)$ may be written as the product of two terms, a rate and a probability. The first is the instantaneous rate of leaving state of origin i irrespective of destination and the second is the conditional probability of selecting j as the destination provided the state of origin is left (i.e. upon leaving i). The first term, the exit rate, determines the timing of the transition while the second, the destination probability, determines the destination (new attribute). The rate of leaving, also known as attrition rate, is

$$\mu_{i+}(x) = \mu_i(x) = \sum_{j \neq i} \mu_{ij}(x)$$

Note that $\mu_i(x) = \mu_{ii}(x)$.

The transition rate is

$$\mu_{ij}(x) = \mu_{i+}(x) \xi_{ij}(x)$$

where $\xi_{ij}(x)$ is the probability that an individual who leaves i selects j as the destination. It is the conditional probability of a *direct transition* from i to j . Note that the above expression is that of a competing risk model or a transition rate model with multiple destinations (Blossfeld and Rohwer, 2002). In the terminology of competing risks, the first term is the rate of event and the second term (destination) indicates the type of event. If the occurrence of the event and the type of event are unrelated, the two terms may be estimated and studied separately (Hachen, 1988, p. 29; Sen and Smith, 1995, p. 372). The first term is studied using a transition rate (hazard rate) model; the second using a logit model or a logistic regression model.

In the migration literature, the first term $\mu_i(x)$ is known as the generation component and the second $\xi_{ij}(x)$ as the distribution component (Rogers et al., 2002).

The discrete-time transition probabilities are related to the probabilities of direct transition in an interesting way. The off-diagonal elements of $\mathbf{M}(x,y)$ may be replaced by $-\mu_{i+}(x,y) \xi_{ij}(x)$ where $\mu_{i+}(x,y)$ is the rate of leaving I , which is assumed to be constant in the interval from x to y . The diagonal elements are $\mu_{i+}(x,y)$. The μ -matrix may be written as

$$\begin{bmatrix} \mu_{11}(x) & -\mu_{21}(x) & \dots & -\mu_{11}(x) \\ -\mu_{12}(x) & \mu_{22}(x) & \dots & -\mu_{12}(x) \\ \dots & \dots & \dots & \dots \\ -\mu_{11}(x) & -\mu_{21}(x) & \dots & \mu_{11}(x) \end{bmatrix} = \begin{bmatrix} \xi_{11}(x) & -\xi_{21}(x) & \dots & -\xi_{11}(x) \\ -\xi_{12}(x) & \xi_{22}(x) & \dots & -\xi_{12}(x) \\ \dots & \dots & \dots & \dots \\ -\xi_{11}(x) & -\xi_{21}(x) & \dots & \xi_{11}(x) \end{bmatrix} \begin{bmatrix} \mu_{1+}(x) & 0 & \dots & 0 \\ 0 & \mu_{2+}(x) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mu_{1+}(x) \end{bmatrix}$$

Now we introduce covariates. As above, they are denoted by Z ($Z = \{Z_1, Z_2, Z_3, \dots\}$). The exit rate is modeled using a transition rate model for a single event (leaving the state of origin). The elementary transition rate model is the basic exponential model, with the rate being independent of age (Blossfeld and Rohwer, 2002):

$$m_i = \exp[\beta_{i0} + \beta_{i1}Z_1 + \beta_{i2}Z_2 + \dots]$$

The model may be written as a log-linear model

$$\ln m_i = \beta_{i0} + \beta_{i1}Z_1 + \beta_{i2}Z_2 + \dots$$

The model is also known as the log-rate model (see e.g. Yamaguchi, 1991, Chapter 4).

The age dependence may be introduced in two ways: non-parametric and parametric. In the first approach, the population is stratified by age and a transition rate is estimated for each age separately. In the parametric approach, age dependence is represented by a model. A common model is the Gompertz model, which imposes onto the transition rate an exponential change with duration. The Gompertz model has two parameters and each may be made dependent on covariates (For detailed treatment, see Blossfeld and Rohwer, 2002). Other parametric models of duration dependence may be used. In studies of marriage and fertility, the Coale-McNeil model is often used to describe the age dependence of the marriage or first birth rate. In migration studies, the model migration schedule is a common representation of the age dependence of the migration rate. Each parameter of the model may be related to covariates. In practice, only one or a selection of parameters is assumed to depend on covariates. TDA (Transition Data Analysis) has a facility for user-defined rate models (Rohwer and Pötter, 1999, Section 6.17.5). The programme may be downloaded from prof. Rohwer's homepage:

<http://www.stat.ruhr-uni-bochum.de/>

The manual (extensive) can be downloaded from the same site. Willekens (2002) has written a brief introduction to TDA with examples.

In some cases, the researcher is not interested in the age dependence of migration rates, but in the effect of covariates on the migration level. Rather than omitting age altogether, as in the basic exponential model, the migration rate is allowed to vary with age but the effect of the covariates on the migration rate does not vary with age. The transition rate model that results is a Cox proportional hazard model. It is written as

$$m_i(x) = m_{i0}(x) \exp[\beta_{i0} + \beta_{i1}Z_1 + \beta_{i2}Z_2 + \dots]$$

where $m_{i0}(x)$ is the baseline hazard. It is the set of age-specific migration rates for the reference category. Note that if the age dependence (age structure) of migration is independent of the dependence on covariates (motivational structure), the baseline hazard

may be represented by a parametric model and the two components may be estimated separately.

This brief discussion illustrates that transition rate models are ideally suited to impose age structures onto migration data. The same applies for spatial structures and motivational structures.

2.4 From transition probabilities to transition rates

In this section, we assume that migration is measured in discrete-time. Examples include the census (based on the residence at time of census and 5 years prior to the census). From that information, the approximate transition rates can be derived. The problem is equivalent to one in which we are given $\mathbf{P}(x,y)$ and $\mathbf{M}(x,y)$ is required. The derivation starts with the exponential expression $\mathbf{P}(x, y) = \exp[-(y - x)\mathbf{M}(x, y)]$. The exponential expression may be approximated by the linear model:

$$\mathbf{P}(x, y) = [\mathbf{I} + \frac{1}{2}\mathbf{M}(x, y)]^{-1} [\mathbf{I} - \frac{1}{2}\mathbf{M}(x, y)]$$

The approximation is adequate when the transition rates are small or the interval is short.

The derivation of the rate of migration during an interval from information in regions of residence at two consecutive points in time is known as the inverse problem: transition rates are derived from transition probabilities (Singer and Silverman, 1979).

$$\mathbf{M}(x, y) = \frac{y-x}{2} [\mathbf{I} - \mathbf{P}(x, y)] [\mathbf{I} + \mathbf{P}(x, y)]^{-1}$$

The inverse relation may be used to infer transition probabilities for intervals that are different from the measurement intervals. For instance, if changes of residence are recorded over a period of five years, the inverse relation may be used to infer the average migration rates $\mathbf{M}(x,y)$ and the transition probabilities over a one-year period. The expression is

$$\mathbf{P}(x, x+1) = \exp[-\mathbf{M}(x, x+1)]$$

where $\mathbf{M}(x,x+1)$ is estimated from $\mathbf{P}(x,y)$ using the inverse method. The method assumes that migration rates are constant during the (x,y) -interval and that the linearity is an adequate approximation of the exponential model.

2.5 Counts

In Section 2.3, the basic transition rate model was related to the log-rate model. The transition rate is the ratio of number of events over total exposure time. These two components may be studied separately, as is done in the log-rate model. In that model, it is assumed that changes in the number and timing of events do not significantly affect the total exposure time. The assumption is realistic when exposure time is large compared to

the number of events. If a variation in number or timing of events does not affect total exposure, the latter component may be considered to be fixed and may be treated as an offset in probability models including regression models. The problem of modeling migration reduces to the prediction of the number of events (counts). The number of events that occur during a unit interval is often represented by a Poisson random variable. The number of events that may occur during the interval is not restricted in any way. Subjects in a (sample) population may experience more than one event during the unit interval. The Poisson model is

$$\Pr\{N_i = n_i\} = \frac{\lambda_i^{n_i}}{n_i!} \exp[-\lambda_i]$$

where N_i denotes the number of migrations originating in i , n_i the observed number of migrations, and λ_i the expected number of migrations. The latter is the parameter of the Poisson model. The parameter may be made dependent on covariates:

$$E[N_i] = \lambda_i = \exp[\beta_{i0} + \beta_{i1}Z_1 + \beta_{i2}Z_2 + \dots]$$

The model may be written as a log-linear model:

$$\ln \lambda_i = \beta_{i0} + \beta_{i1}Z_1 + \beta_{i2}Z_2 + \dots$$

In principle, Z_p can be any covariate. In conventional log-linear analysis, all covariates are discrete or categorical. The observations on event occurrences may therefore be arranged in a contingency table. The covariates refer to rows, columns, layers and combinations of these (to represent interaction effects). Log-linear models of age and spatial structures of migration flows are studied by Rogers et al. (2003).

The log-rate model is a log-linear model with an offset:

$$E\left[\frac{N_i}{PY_i}\right] = \frac{\lambda_i}{PY_i} = \exp[\beta_{i0} + \beta_{i1}Z_1 + \beta_{i2}Z_2 + \dots]$$

Since PY_i is fixed, the equation may be rewritten as follows:

$$E[N_i] = \lambda_i = PY_i \exp[\beta_{i0} + \beta_{i1}Z_1 + \beta_{i2}Z_2 + \dots]$$

3. Incomplete data

In the previous section, it is assumed that the data are adequate to estimate the parameters of the probability models that are specified. The method applied is the maximum likelihood method. In this section the assumption is relaxed. Some data may be missing.

If data are missing, the strategy consists of two steps. The first is to predict the missing data and second step is to estimate the parameters of the model assuming that the data are complete. This two-step procedure is the EM- algorithm (McLachlan and Krishnan,

1997). Suppose we are interested in migration by origin and destination (N_{ij}), but the data are limited to departures and arrivals by region (n_{i+} and n_{+j}). The model is

$$E[N_{ij}] = \lambda_{ij} = \alpha_i \beta_j$$

In the first step, the expected value of N_{ij} is determined assuming values for the parameters (*expectation*). Assume $\alpha_i=1$ and $\beta_j=1$. Hence $E[N_{ij}] = 1$

In the second step, the parameters of the model are estimated by maximizing the probability that the model predicts the data. If the observations are independent, the probability model is

$$\Pr\{N_{ij} = n_{ij}\} = \frac{\lambda_{ij}^{n_{ij}}}{n_{ij}!} \exp[-\lambda_{ij}]$$

$$\text{with } E[N_{ij}] = \lambda_{ij} = \alpha_i \beta_j$$

The *maximization* of the probability is equivalent to maximizing the log-likelihood:

$$l = \sum_{ij} [n_{ij} \ln[\alpha_i \beta_j] - \alpha_i \beta_j]$$

The first-order conditions result in the following equations:

$$\hat{\alpha}_i = \frac{n_{i+}}{\sum_j \hat{\beta}_j} \text{ and } \hat{\beta}_j = \frac{n_{+j}}{\sum_i \hat{\alpha}_i}$$

The EM algorithm results in the well-known expression

$$\lambda_{ij} = \frac{n_{i+} n_{+j}}{n_{++}}$$

A similar procedure is applied when initial guestimates are available (m_{0ij}) (Willekens, 1999).

4. Conclusion

The paper presents a unified perspective on the modeling of migration. The perspective is derived from event history analysis and multistate modeling. The level and direction of migration may be represented by different data types. Three types are distinguished: counts, rates and probabilities. The different representations of migration may be reduced to one of these types or a combination. Different types of migration data may be modeled using models that have been developed for the analysis of life history data. The core consists of a (multistate) transition model in discrete time and continuous time. Transitions recorded in continuous time are direct transitions or events (migrations). The model of migration is a transition rate model. Transitions recorded in discrete time refer to migrants who are identified by comparing the place of residence of a person at two

consecutive points in time. The appropriate model is a transition probability model which is a multinomial logistic regression model.

A major conclusion of the paper is that migration should be treated as a life event. Models that have been developed to study life events and life histories are perfectly suited for the study of migration. They allow the analysis of data of different types and the conversion of one data type into another. They also allow the treatment of incomplete data. Initially migration models were spatial interaction models. Today they are increasingly considered as applications of event history models.

References

- Andersen, P.J., O. Borgan, R.D. Gill and N. Keiding (1993) Statistical models based on counting processes. Springer Verlag, New York.
- Andersen, P.K. and N. Keiding (1996) Survival analysis. In: P. Armitage and H.A. David eds. Advances in biometry. Wiley, New York, pp. 177-199.
- Aoki, M. (1976) Optimal control and system theory in dynamic economic analysis. North Holland, New York.
- Blossfeld, H.P. and G. Rohwer (2002) Techniques of event history modeling. New approaches to causal analysis. Lawrence Erlbaum, Mahwah, New Jersey. Second edition.
- Çınlar, E. (1975) Introduction to stochastic processes, Prentice-Hall, Englewood Cliffs, New Jersey.
- Courgeau, D. (1973) Migrants et migrations. Population, 28, pp. 95-128.
- Courgeau, D. (1982) Comparaison des migrations internes en France et aux Etats-Unis, Population, 6, pp. 1184-1188.
- Director, S.W. and R.A. Rohrer (1972) Introduction to system theory. McGraw-Hill, New York.
- Hachen, D.S. (1988) The competing risk model. *Sociological Methods and Research*, 17(1):21-54. Reprinted in in D.J. Bogue, E.E. Arriaga and D.L. Anderton eds. Readings in population research methodology. Social Development Center, Chicago, and UNFPA, New York, pp. 21.85-21.101.
- Hosmer, D.W. and S. Lemeshow (1999) Applied survival analysis. Regression modeling of time to event data. Wiley, New York.

Kitsul, P. and D. Philipov (1981) The one year/five year migration problem; in: A. Rogers ed. *Advances in multiregional demography*, Research Report RR-81-6, IIASA, Laxenburg, Austria, pp. 1-33.

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

Long, J.F. and C.G. Boertlein (1981) Using migration measures having different intervals, manuscript. U.S. Bureau of the Census, Washington D.C.

McLachlan, G.J. and T. Krishnan (1997) *The EM algorithm and extensions*. Wiley, New York.

Rajulton, F. (1999) LIFEHIST: Analysis of life histories: a state-space approach. Paper presented at the Workshop on Longitudinal Research in Social Science: A Canadian Focus, Windermere Manor, London, Ontario, Canada, October 25-27, 1999.

Rohwer, G. and U. Pötter (1999) TDA User's manual. Ruhr-Universität Bochum. Fakultät für Sozialwissenschaften, Bochum, Germany.

Rogers, A. (1975) *Introduction to multiregional mathematical demography*. Wiley, New York.

Rogers, A., F. Willekens, K. Little and J. Raymer (2002) Describing migration spatial structure. *Papers in Regional Science: Journal of the Regional Science Association International*, 81(1):29-48

Rogers, A., F. Willekens and J. Raymer (2003) Imposing age and spatial structures on inadequate migration flow data sets. *The Professional Geographer*, 55(1):56-69

Schoen, R. (1988) *Modeling multigroup populations*. Plenum Press, New York.

Sen, A. and T. Smith (1995) *Gravity models of spatial interaction behavior*. Springer Verlag, Berlin.

Singer, B. and S. Spilerman (1979) Mathematical representations of development theories. In: J.R. Nesselroade and P.B. Baltes eds. *Longitudinal research in the study of behavior and development*. Academic Press, New York, pp. 155-177.

Strang, G. (1980) *Linear algebra and its applications*. Second edition. Academic Press, New York.

Strauss, D. and R. Shavelle (1998) An extended Kaplan-Meier estimator and its applications. *Statistics in Medicine*, 17:971-982

Therneau, T.M. and S.M. Grambsch (2000) Modeling survival data. Extending the Cox model. Springer Verlag, New York.

Willekens, F.J. (1999) Modeling approaches to the indirect estimation of migration flows: from entropy to EM. *Mathematical Population Studies*, 7(3):239-278

Willekens, F.J. (2002) An introduction to TDA. Manuscript.

Yamaguchi, K. (1991) Event history analysis. Sage Publications, Newbury Park, USA.