# Models for Grouped Transition Data

**by**

**Reinhard Hujer, Kai-Oliver Maurer and Marc Wellner**

**Department of Economics**
**Johann Wolfgang Goethe-University, Frankfurt am Main**

**March 1996**

Financial support from the Deutsche Forschungsgemeinschaft (#Hu 225/7-1; "Weiterbildung und Erwerbsverläufe") is gratefully acknowledged.

Reinhard Hujer, Kai-Oliver Maurer and Marc Wellner
Department of Economics
Johann Wolfgang Goethe-University
Mertonstraße 17
D-60054 Frankfurt am Main

Phone: +49 (0)69 798-28115
Fax: +49 (0)69 798-23673

E-Mail: hujer@wiwi.uni-frankfurt.de/K.Maurer@em.uni-frankfurt.de/Wellner@em.uni-frankfurt.de

WWW: http://www.wiwi.uni-frankfurt.de/professoren/empwifo/

# ABSTRACT

This paper is intended as a short survey of the most relevant methods for grouped transition data. The fundamentals of duration analysis are discussed in a continuous time framework, whereas the treatment of methods for discrete durations is limited to the peculiarity of these models. In addition, some recent empirical applications of the methods are discussed.

# I. Introduction

Hazard models – or models for transition data, as LANCASTER (1990) calls them to stress the fact that the econometrician is usually not only interested in the duration of an event but also in the destination that is entered at its end – are a rapidly growing field especially in empirical labour market analysis. Within these methods, models that explicitly account for the fact that economic data generally are either rounded, grouped or collected at fixed intervals have been getting more attention in empirical applications in the last few years.

This paper is the result of our studies concerning the theory of hazard models, which we have done as the first step in our current research project regarding the effects of training on employment histories of individuals both in the western and eastern part of reunified Germany. The aim of the paper is to summarise the relevant methods as developed in econometric theory and practice, thereby giving an overview of the theory and the empirical literature on grouped duration models. The fundamentals of duration analysis will be demonstrated within the framework of continuous time, whereas the treatment of discrete data is limited to the peculiarity of these models.

The paper is organised as follows: In Chapter II we present the theoretical models for transition data. Recent and – with regard to the methods presented in chapter II – important empirical applications of grouped duration models in the field of labour econometrics are discussed in chapter III. A short conclusion completes this paper.

# II. Theoretical Models for Grouped Transition Data

## II.1 Cox's Proportional Hazards Model

The starting point for the derivation of models for grouped transition data are continuous hazard models. They are called continuous, because duration $T_i$ of observation i is said to be a continuous random variable.[1] In this and the next two sections we will use continuous time models to introduce the most important concepts of transition models, before we are looking at the notion of discrete time (or grouped data).

A well known and widely used model for continuous transition data is the Proportional Hazards model as proposed by COX (1972):[2]

$$\lambda_i\left(t|x_i\right)=\Phi\left(x_i,\beta\right)\lambda_0(t) \qquad (1)$$

$\lambda_0$ is called the *baseline hazard*, because it corresponds to $\Phi(\cdot)=1$, $x_i$ is a vector of exogenous variables, $\beta$ is the vector of the coefficients to be estimated, t is a realisation of $T_i$ and $\lambda_i\left(t|x_i\right)=\lim\limits_{dt\to 0+}P\left(t\leq T_i<t+dt|t\leq T,x_i\right)\cdot(dt)^{-1}$ is the hazard rate, i.e. the instantaneous rate of leaving a certain state of interest per unit time period at t (LANCASTER (1990)). The probability of survival to t is given by the corresponding *survivor function* $S_i\left(t|x_i\right)=\exp-\int_0^t\lambda_i(u)du$. Durations refer to the times from the beginning of spells.

In contrast to parametric models, the Proportional Hazards Model allows a far more flexible approach, because the form of the baseline hazard does not need any further specification. Instead, $\lambda_0$ is to be estimated. Thus, the Proportional Hazards model avoids an assumption regarding the distribution of duration $T_i$ and thus does not imply a particular evolution of hazard

[1] Generally, observations may either be spells or individuals. In a single spell framework, however, this distinction is redundant.

[2] The name of the model derives from the fact that the hazards for two individuals with vectors of covariates $x_1$ and $x_2$ are in the same ratio for all t (LANCASTER (1990)), which is a quite strong assumption. This property of the model vanishes if individual covariates are allowed to vary over time.

and survivor functions over time (*duration dependence*) as parametric hazard models do. For that reason, it is also termed as *semiparametric*.

Usually $\Phi$ is chosen as $\exp(x_i'\beta)$. This specification of $\Phi$ is convenient, because, as $\lambda_i(t|x_i)$ has to be nonnegative, no restrictions need to be imposed on $\beta$. Furthermore, as can easily be shown, it admits a convenient interpretation of the Proportional Hazards Model as a linear model with $\varepsilon_i$ distributed as a Type-I extreme value random variable (KIEFER (1988A), RIDDER (1990)):[3]

$$\ln \int_0^{T_i=t} \lambda_0(u)\,du = t_i^* = -x_i'\beta + \varepsilon_i \qquad (2)$$

The expression on the far left hand side is the logarithm of the *integrated baseline hazard* $\Lambda_0(t)$. Since it could be conceived as a transformed duration $t_i^*$, formulation (2) is also called a *Generalized Accelerated Failure-Time* (GAFT) model (RIDDER (1990)). As KIEFER (1988A) points out, least squares estimation methods can only be used for this linear specification, if the data are not heavily censored and a correction for the estimate of the intercept is made in order to account for the nonzero mean of $\varepsilon_i$. However, this requires knowledge of the integrated baseline hazard.

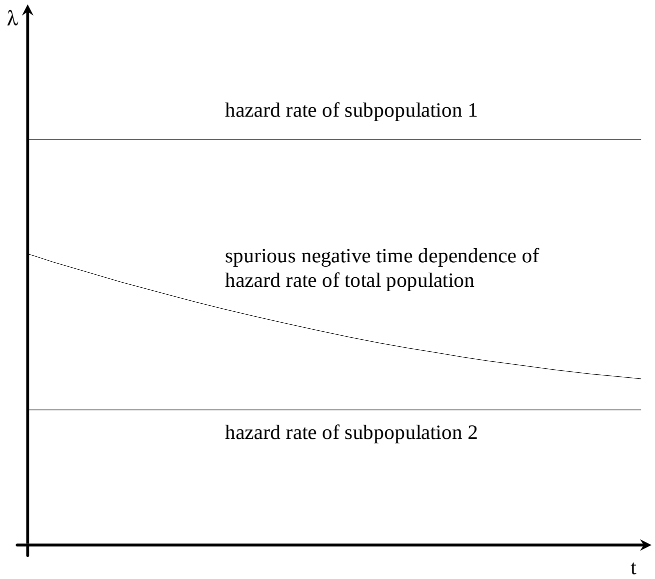
## II.2 Spurious Time Dependence

The observation units in the data set are different from each other for many reasons – there could be personal differences, variances in the surroundings, etc. Ideally, the covariates in the specification (1) allow for all potential differences between observation units. In practice, however, this is rather unlikely. Reasons might be shortcomings in the data set or neglect of the econometrician, but usually it is simply impossible to include every possible source of variation in the regressors. The part of heterogeneity between the individuals that is not explicitly accounted for in the vector of covariates is usually called *unmeasured heterogeneity*. Therefore, an additional term, the hererogeneity term, should be included in the hazard to account for this source of variation.

[3] Note that $-\varepsilon_i$ is also distributed as an Type I-extreme value random variable (JOHNSON/KOTZ (1970), p.272).

Failure to control for unmeasured heterogeneity can result in downward biased estimators and *spurious time dependence* (ELBERS/RIDDER (1982), KIEFER (1988A)): It is possible that the estimated hazard function declines more steeply (or rises more slowly) than the true hazard (or declines instead of rising or being constant). Intuitively, in a simple model, where the underlying population consists of two different subpopulations, each with a distinct and constant hazard rate, individuals of the subpopulation with the higher transition rate tend to leave the state of interest sooner. Thus, the portion of individuals from the subpopulation with the lower transition rate in the population at risk increases and the transition rate for the whole population seems to decrease over time, if unmeasured heterogeneity is not accounted for (figure 1).

*Figure 1:*
*Spurious time dependence as a result of neglecting unmeasured heterogeneity*



In the context of the Proportional Hazards model the heterogeneity term $\theta_i = \exp(v_i)$ typically enters the hazard multiplicatively. The resulting model is also known as the *Mixed Proportional Hazards model* (RIDDER (1990)):

$$\lambda_i\left(t \middle| x_i, v_i\right) = \lambda_0(t)\exp\left(x_i'\beta + v_i\right) \qquad (3),$$

where $v_i$ is an individual specific random variable with $E\left[v_i\right] = 0$, i.e. $E\left[\exp\left(v_i\right)\right] = 1$. The Mixed Proportional Hazards model can also be written as a Generalized Accelerated Failure Time model:

$$\ln\int_0^{T_i=t}\lambda_0(u)\,du = t_i^* = -x_i'\beta+\varepsilon_i = -x_i'\beta+\eta_i+\kappa_i \qquad (4).$$

Now, $\eta_i$ is a type-I extreme value random variable representing the Proportional Hazards portion of the specification, $\kappa_i = -v_i$ represents the unobserved heterogeneity, and the distribution of $\varepsilon_i$ is a convolution of the distributions of $\eta_i$ and $\kappa_i$ (SUEYOSHI (1994)).

The fundamental problem of the integration of unobserved heterogeneity is the fact that the distribution of $v_i$ ($\kappa_i$), the so called *mixing distribution*, is, of course, unknown. In econometric literature, two solutions to overcome this problem have been proposed:

(1) One may simply assume a particular shape for the distribution of the heterogeneity term. Then, the relevant parameters of that distribution are also to be estimated. Often the Gamma distribution is chosen for computational simplicity as it gives a closed form expression for the likelihood, avoiding integration by numerical methods (LANCASTER (1979), TUMA/HANNAN (1984), MEYER (1987), HUJER/SCHNEIDER (1996)).

(2) In a widely cited series of papers, HECKMAN/SINGER (1982, 1984A, 1984B, 1985, 1986) instead propose nonparametric methods to assess the distribution, because of sensitivity of parameter estimates to the assumed shape of the distribution of the heterogeneity term. As they argue, specification of both the functional form of structural duration distributions and the functional form of the distribution of unobservables leads to overparameterisation of the model. One should rather approximate the true distribution of the heterogeneity term by fitting a discrete distribution with a finite number of mass points. This method is used, for instance, by HUJER/SCHNEIDER (1989) or NARENDRANATHAN/STEWART (1993B).

As TRUSSELL/RICHARDS (1985) point out, much of the parameter instability found by HECKMAN/SINGER (1982, 1984A, 1984B, 1985, 1986) might be the result of their assumption of a Weibull baseline hazard. Therefore, when estimating $\lambda_0(t)$ nonparametrically, the functional form of the heterogeneity distribution may as well be unimportant. MEYER (1987, 1990) takes the same view, although he does not prove his assertion. Another problem of the Heckman-Singer approach is the fact that number of mass points is not predetermined but is to be assessed using an iterative procedure (HECKMAN/SINGER (1984A), TRUSSELL/RICHARDS (1985)). NARENDRANATHAN/STEWART (1993B) compare a two mass point mixing model using the Heckman-Singer procedure with a normal mixture model and get very similar results.

## II.3 Competing Risks

The discussion so far has been restricted to *single risk models*. In these models there is only one kind of event terminating the duration spent in the state of interest. Most studies regarding the effect of unemployment compensation on the duration of unemployment, for example, focus on the exit of unemployment through transition into employment. In reality, however, further causes (or *risks*) for leaving the state of interest can exist. Currently unemployed may not only find a new job, but alternatively may leave the unemployment register by completely dropping out of the workforce (exit into non-employment). It is then appropriate to extend the framework to *competing risks*.

Let us assume R possible mutually exclusive destination states. Note, that any set of destination states always can be redefined as to yield mutually exclusivity. Consider independent *latent* random variables $T_{ir}^*,\ r = 1, \dots, R$, each measuring the time until event of type r (transition in state r). The duration $T_i$ we observe in practice is the minimum of these theoretical durations:

$$T_i = \min_r T_{ir}^*$$

One can now define R *(cause-specific) transition intensities*:

$$\lambda_{ir}\left(t \middle| x_{ir}\right) = \lim_{dt \to 0+} P\left(t \le T_i < t + dt, Y_i = r \middle| t \le T_i, x_{ir}\right) \cdot (dt)^{-1} \qquad \forall\, r \qquad (5).$$

$Y_i$ is a variable indicating which of the R events occurs. $\lambda_{ir}\left(t|x_{ir}\right)$ represents the probability of transition to state r after duration time t, conditional on not having left prior to t in presence of the other possible events and on the set of cause-specific covariates $x_{ir}$. In the context of the Mixed Proportional Hazards model, the $\lambda_{ir}\left(t \middle| x_i\right)$ can be written as:

$$\lambda_{ir}\left(t \middle| x_{ir}, v_{ir}\right) = \lambda_{0r}(t) \exp\left(x_{ir}'\beta_{ir} + v_{ir}\right) \qquad \forall\, r \qquad (6).$$

The *(overall) hazard function* gives the instantaneous rate for failure of any type. As the R destination states are mutually exclusive, it equals the sum of all cause-specific transition intensities:

$$\lambda_i\left(t \middle| x_{i1},\ldots,x_{iR}\right) = \sum_r \lambda_{ir}\left(t \middle| x_{ir}\right) \qquad (7).$$

In the case of competing risks inclusion of the heterogeneity component in the transition intensities as in (6) raises the question whether it can be safely assumed that these disturbances are independent across the intensities (NARENDRANATHAN/STEWART (1993A)). Only in that case can the transition intensities be estimated one at a time as in the single risk case. Dependent competing risks, i.e. risks with correlations across the individual transition intensities, will be discussed in section II.4.4.

## II.4 Discrete Hazard Rate Models

### II.4.1 Discrete Hazard Rate Models and Panel Data

Until now, we have been restricting our discussion to the concept of continuous time to introduce the most important concepts in duration analysis. The use of continuous time models in econometric practice is only justified, if the times, at which the events of interest occur, can be exactly determined. However, adequate data are usually not accessible. In a narrow sense, all economic data are not available on a continuous time basis. Some data are quite near to the ideal (as are stock exchange prices for example), but especially microeconometric databases like panels are based on weekly, monthly or even yearly interviews and, as a result, only time intervals can be specified, in which certain events have occurred. Economic data are also frequently rounded or grouped. In the literature of duration analysis all three phenomenons (interval spacing, rounding, grouping) are subsumed under the term *grouped data* (KIEFER (1988B)).

When using continuous time models with grouped duration data, problems result from the existence of *ties*, equal durations for different observations. In continuous data, true ties are the exception, but with grouped data, many ties can be expected. As a consequence, the parameter estimates of various models (as of the Cox model for instance) are useless (BLOSSFELD/HAMERLE/MAYER (1986); see also KALBFLEISCH/PRENTICE (1980), COX/OAKES (1984)). Often the application of discrete hazard rate models is called for in this case, though the issue is not yet settled (SUEYOSHI (1995)): HECKMAN/SINGER (1984B) or LANCASTER (1990) instead propose to work in continuous time and translate to discrete time as necessary. In empirical studies, MEYER (1990), for instance, applies discrete time and GRITZ (1993) continuous time models to weekly data. In the end, the discrete methods presented below are nothing more than a mapping from a continuous-time specification to the discrete observations (SUEYOSHI (1995)).

The data which we will be using are panel data from the Socio-Economic Panel (SOEP) for Germany. Individuals in this panel are interviewed once a year, but data for the individuals' spells are available on a monthly basis. Of course, in such a huge database, ties are a common feature in this context, especially since one typically observes a kind of "year-bias" in duration data: first, in retrospective interviews people often are mistaken about the correct duration of a certain spell and tend to specify longer durations in entire years even if a more accurate time scale is at hand (memory bias); secondly, many contracts, benefits, etc. usually expire after 12, 18, 24 or so months. Thus, the application of discrete models is necessary for our project.

### II.4.2 Specification of Discrete Hazard Rate Models

We consider the case where individual duration data are grouped into $J+1$ intervals with the j-th interval defined as $\left[t_j, t_{j+1}\right)$, $j = 0,1,\ldots J$. For an arbitrary j the discrete hazard rate $h_i\left(j \middle| x_i\right)$ is defined as the probability that a spell ends before $t_{j+1}$, given that it has lasted at least until $t_j$ and the set of covariates:

$$h_i\left(j \middle| x_i\right) = P\left[T_i < t_{j+1} \middle| T_i \geq t_j,\, x_i\right] = 1 - S_i\left(t_{j+1} \middle| x_i\right) \cdot S_i\left(t_j \middle| x_i\right)^{-1} \qquad (8).$$

Specifying the continuous hazard function $\lambda_i(t)$ that corresponds to $S_i\left(t \middle| x_i\right)$ using the Proportional Hazards model (1) and substituting $\gamma(j) = \ln \int_{t_j}^{t_{j+1}} \lambda_0(u)du$ gives the following expression for the hazard rate:

$$h_i\left(j \middle| x_i\left(t_j\right)\right) = 1 - \exp\left[-\exp\left(x_i'\left(t_j\right)\beta + \gamma(j)\right)\right] \qquad (9).$$

Notice that we have also introduced time varying covariates $x_i\left(t_j\right)$. Specification (9) requires that the covariates are constant over the interval $\left[t_j, t_{j+1}\right)$ (NARENDRANATHAN/STEWART (1993B)).

Also note that the discrete hazard rate has the shape of a Type I-extreme value distribution even without any distributional assumptions regarding any of the variables of the model so far. Thus, the variable $\varepsilon_i\left(t_j\right) = x_i'\left(t_j\right)\beta + \gamma\left(t_j\right)$ is – as in the GAFT-specification (2) for continuous time – again a Type I-extreme value random variable. HAN/HAUSMAN (1990) use this fact and the relation of the model above to ordered response models (MADDALA (1983)) to specify the likelihood function over the density of $\varepsilon_i\left(t_j\right)$.

Alternatively, the likelihood function can be derived in the following way: Define a dummy variable $\delta_i$, indicating whether the observation of the i-th individual is right-censored ($\delta_i = 0$) or not. $k_i$ is either the interval, in which an event for individual i can be observed ($\delta_i = 1$), or the censoring interval ($\delta_i = 0$). For a sample of N individuals the likelihood function is (MEYER (1987, 1990)):

$$L(\gamma,\beta) = \prod_{i=1}^{N}\left[\left[\left\{\underbrace{1 - \exp\left(-\exp\left[x_i'\left(t_{k_i}\right)\beta + \gamma\left(k_i\right)\right]\right)}_{(1)}\right\}\right]^{\delta_i} \cdot \underbrace{\prod_{m=0}^{k_i-1}\exp\left\{-\exp\left[x_i'\left(t_m\right)\beta + \gamma(m)\right]\right\}}_{(2)}\right] \quad (10).$$

Part (1) of (10) is the hazard rate (9), the probability of having an event in interval $k_i$ conditional on survival to that interval. The whole part between both product signs equals one except when a spell ends in interval $k_i$. Part (2) is the probability of survival at least until $t_{k_i}$, the *overall survivor function* $S_i\left(t_{k_i}, x_i, \beta\right)$. Survival to $t_{k_i}$ is the same as surviving each of the preceding intervals $\left[t_m, t_{m+1}\right)$ for $m = 0,\ldots,k_i - 1$, so the overall survivor function may be expressed in terms of interval specific, conditional survivor functions $\alpha$ (KIEFER (1988B), SUEYOSHI (1995)):

$$S_i\left(t_{k_i}, x_i, \beta\right) = \prod_{m=0}^{k_i-1}\alpha_{im}\left(x_i\left(t_m\right), \beta\right) \qquad (11a)$$

where

$$\begin{aligned}\alpha_{im}\left(x_i\left(t_m\right),\beta\right) &= S_i\left(t_{m+1}, x_i, \beta \middle| T_i \geq t_m\right) \\ &= \exp\left(-\int_{t_m}^{t_{m+1}}\lambda_i\left(s, x_i\left(t_m\right),\beta\right)ds\right) = 1 - h_i\left(m \middle| x_i\left(t_m\right)\right)\end{aligned} \qquad (11b).$$

Thus, the probability that a spell ends in interval $k_i$, which is the probability of surviving the first $k_i - 1$ intervals but not the $k_i$-th is given by (SUEYOSHI (1995)):

$$P\left(t_{k_i} \leq T_i < t_{k_i+1}\right) = S_i\left(t_{k_i}, x_i, \beta\right) - S_i\left(t_{k_i+1}, x_i, \beta\right) = \left(1 - \alpha_{ik_i}\left(x_i\left(t_{k_i}\right), \beta\right)\right) \cdot \prod_{m=0}^{k_i-1} \alpha_{im}\left(x_i\left(t_m\right), \beta\right) \quad (12a)$$

$$= h_i\left(k_i \middle| x\left(t_{k_i}\right)\right) \cdot \prod_{m=0}^{k_i-1} \alpha_{im}\left(x_i\left(t_m\right), \beta\right) \quad (12b).$$

Formulation (12b) leads us to the above likelihood function. A fundamental assumption underlying the derivation of the likelihood (10) and of the overall survivor function (11a) is that censoring occurs at the beginning of intervals (KIEFER (1990)). Thus, $k_i$ is the first interval of the following new spell in case of a transition or the first interval in which the individual is not observable anymore in case of right-censoring.

Discrete hazard models have an intriguing relationship to binomial models that allows us to reformulate the likelihood function (10) in order to get a more familiar specification. Instead of the each individual or each spell, each individual-period combination may be conceived as a separate observation: Each individual contributes $k_i$ observations, one for each interval j he enters. This leads to a sample size of $N^* = \sum_i k_i$ observations. Then, a new dummy variable $d_n$ can be defined, taking a value of zero, if the spell was completed in the n-th individual-interval, and a value of one, if the n-th individual-interval was survived. An individual i thus contributes for a completed spell with a duration of six intervals, for instance, a sequence of six observations: 0, 0, 0, 0, 0, 1. The likelihood function can then be written as (KIEFER (1988B)):

$$L^*(\gamma, \beta) = \prod_{n=1}^{N^*} \alpha_n^{\,d_n} \left(1 - \alpha_n\right)^{1-d_n} \quad (13).$$

This specification is similar to the standard binary response likelihood. The only difference is that the usual normal or logistic cumulative distribution functions are replaced by the interval specific survivor functions depending upon integrated hazards. It is tempting, however, to simply disregard this difference and estimate a common binary (or with competing risks a multinomial) logit or probit model. As we will see below, this is often done in empirical applications. However, as SUEYOSHI (1995) points out, one should consider the implications for hazard behaviour of different specifications for the $\alpha_n$. The results of his comparison of various specifications for the $\alpha_n$ indicate very different effects of changes in the explanatory variables.

The probit model also tends to depart quite far from propotionality as opposed to logistic or extreme value specifications.

### II.4.3 Integration of Competing Risks and Unobserved Heterogeneity

As HUJER/SCHNEIDER (1996, p.58) point out, the best basis for formulating the likelihood function is now the concept of the survivor function. The reason is that the inclusion of an observation-specific heterogeneity term, be it cause-specific or not, does not permit a interval specific factorization – as e.g. in (10) – any longer. Also, when formulating the transition rates it now has to be ensured that no other transition in the same interval occurs. This was unnecessary in the continuous time framework as the probability that more than one transition occurs at the same time is zero if time is a continuous random variable.

Recalling transition intensity (6) and hazard function (7) for continuous time, we obtain for the continuous time overall survivor function for $r = 1, \dots, R$ independent destination states

$$\begin{aligned} S_i\left(t \middle| x_{i1}(t), \dots, x_{iR}(t)\right) &= \exp\left(-\sum_{r=1}^{R} \int_0^t \lambda_{ir}\left(s \middle| x_{ir}(s), \beta_r, v_{ir}\right) ds\right) \\ &= \prod_{r=1}^{R} \exp\left(-\int_0^t \lambda_{ir}\left(s \middle| x_{ir}(s), \beta_r, v_{ir}\right) ds\right) \end{aligned} \qquad (14).$$

The discrete time overall survivor function then is

$$S_i\left(t_j \middle| x_{i1}\left(t_j\right), \dots, x_{iR}\left(t_j\right), v_{i1}, \dots, v_{iR}\right) = \prod_{r=1}^{R} S_{ir}\left(t_j \middle| x_{ir}\left(t_j\right), v_{ir}\right) \qquad (15),$$

where

$$\begin{aligned} S_{ir}\left(t_j \middle| x_{ir}\left(t_j\right), v_{ir}\right) &= \exp\left\{-\sum_{m=0}^{j-1} \int_{t_m}^{t_{m+1}} \lambda_0(s) \exp\left(x_{ir}'(s)\beta_r + v_{ir}\right) ds\right\} \\ &= \exp\left\{-\exp\left(v_{ir}\right) \sum_{m=0}^{j-1} \exp\left(x_{ir}'\left(t_m\right)\beta_r + \gamma_r(m)\right)\right\} \end{aligned} \qquad (16).$$

In addition, we assume that heterogeneity components for different destination states are stochastically independent from each other (HUJER/SCHNEIDER (1996), p.60): $\operatorname{cov}\left(v_{iq}, v_{ir}\right) = 0$ $\forall q \neq r$. If $G\left(\theta_{ir}\right)$ is the unknown mixing distribution function for $\theta_{ir} = \exp(v_{ir})$ equation (16) can be rewritten as:

$$S_{ir}\left(t_j \left| x_{ir}\left(t_j\right), v_{ir}\right.\right) = \int_0^{\infty} \exp\left\{-\theta_{ir} \sum_{m=0}^{j-1} \exp\left(x'_{ir}\left(t_m\right)\beta_r + \gamma_r\left(t_m\right)\right)\right\} dG\left(\theta_{ir}\right) \qquad (17).$$

The likelihood function again can be specified with the help of a dummy variable $\delta_{ir}$ which indicates, if individual i exits in destination state r ($\delta_{ir} = 1$) or not. $k_i$ indicates either the interval of transition or the censoring interval:

$$L = \prod_{i=1}^{N} \prod_{r=1}^{R} \left[\frac{S_{ir}\left(t_{k_i} \left| x_{ir}\left(t_{k_i}\right), v_{ir}\right.\right) - S_{ir}\left(t_{k_i+1} \left| x_{ir}\left(t_{k_i+1}\right), v_{ir}\right.\right)}{S_{ir}\left(t_{k_i+1} \left| x_{ir}\left(t_{k_i+1}\right), v_{ir}\right.\right)}\right]^{\delta_{ir}} S_{ir}\left(t_{k_i} \left| x_{ir}\left(t_{k_i}\right), v_{ir}\right.\right) \qquad (18).$$

### II.4.4 Dependent Competing Risks

So far, in the context of competing risks, it has been assumed that the latent failure times $T^*_{ir}$, $r = 1, \ldots, R$, are independent from each other. This assumption is usually questionable in economic problems. As LANCASTER (1990, p.107) points out, "eliminating a possible destination will generally alter people’s behaviour". If one allows for interdependence between the different risks, simultaneous estimation of the various transitions will be necessary. The intriguing relation of hazard models to the linear regression models via the GAFT-specification is very useful in this respect, because, in linear models, simultaneity can be more easily dealt with than in the direct hazard specification. This fact has been utilised by HAN/HAUSMAN (1990) or BELZIL (1995), for example. To repeat the GAFT-model (4):

$$\ln \int_0^{T_i=t} \lambda_0(u)\, du = t_i^* = -x_i'\beta + \varepsilon_i = -x_i'\beta + \eta_i + \kappa_i \qquad (19).$$

Remember that the distribution of $\varepsilon_i$ ultimately depends on the distribution of the heterogeneity component $\kappa_i$ and that $\eta_i$ is extreme value distributed. HAN/HAUSMAN (1990) simply assume $\varepsilon_i$

to be normally distributed. As the choice of a Gamma distribution for $\kappa_i$ is also arbitrary, this assumption might be equally justified.

Transformed duration $t_i^*$ can be conceived as a latent variable, whereas the number $j \in \{0, \ldots, J\}$ of the interval where transition occurs is observable. One can interpret the spell duration of individual i as a choice between ordered categories, i.e. the discrete intervals. Thus, there is also a close relationship between hazard models and ordered response models, which is quite useful, because interdependence of the competing risks and simultaneous estimation can easily be dealt with in a multivariate ordered response models, especially if $\varepsilon_i$ is assumed to be distributed as a normal or logistic random variable, leading to ordered probit and logit models, respectively. In an ordered response framework based on discrete hazard models, HAMERLE/TUTZ (1989) propose to interpret the latent variable of the model as the sum of forces prolonging survival time or, alternatively, as time-continuous survival time in the respective interval.

# III. Empirical Applications of Grouped Duration Models

This chapter reviews some papers using methods for grouped duration data as discussed in the previous chapter. This overview is not meant to be complete. Instead we focus on those papers that are, in our opinion, most relevant to our project and represent major lines of development in methods for grouped duration data.

**MEYER (1990)**, based on his own earlier work (MEYER (1987)), examines the effects of unemployment insurance benefits on unemployment durations using data from Continuous Wage and Benefit History Unemployment Insurance administrative records for 1978-1983. His sample consists of 3,365 males from twelve U.S. states. The data are available on weekly basis, and MEYER (1990) decides to use duration models for discrete time. The analysis is limited to the transition from unemployment to employment, so that his initial likelihood function is identical to equation (10) in the previous chapter. Unobserved heterogeneity is accounted for by adding a heterogeneity component following a gamma distribution with mean one and variance $\sigma^2$, leading to the following log-likelihood:

$$\begin{aligned} l\left(\gamma,\beta,\sigma^2\right) = \sum_{i=1}^{N} \ln\Bigg\{ & \left[1+\sigma^2\cdot\sum_{m=0}^{k_i-1}\exp\{\gamma(m)+x_i'(m)\beta\}\right]^{-\sigma^{-2}} \\ & -\delta_i\left[1+\sigma^2\cdot\sum_{m=0}^{k_i}\exp\{\gamma(m)+x_i'(m)\beta\}\right]^{-\sigma^{-2}}\Bigg\} \end{aligned} \qquad (20).$$

MEYER (1990) rejects nonparametric methods for measuring unobserved heterogeneity on the basis of the criticism by TRUSSELL/RICHARDS (1985), which has already been outlined above. He compares different specifications without heterogeneity, with gamma distributed heterogeneity and with a Weibull baseline hazard and with non-parametric estimation of the baseline hazard. He finds that the non-parametric estimation of the baseline hazard dominates the Weibull specification, thereby confirming presumptions about the effects of misspecifying the baseline hazard. Surprisingly, however, most estimates are not significantly affected when introducing unobserved heterogeneity. In substance, MEYER'S (1990) results suggest a relatively large disincentive effect of unemployment insurance benefits.

The effect of unemployment insurance on the probability of an individual leaving unemployment is also studied by **NARENDRANATHAN/STEWART (1993B)**, but they give special attention to the variation of this effect over the duration of the individual' s unemployment spell. The underlying data are taken from the UK Department of Health and Social Security Cohort Study of the Unemployed 1978/1979 and seems to be grouped in weekly intervals. The sample consists of 1,571 men, who reported to be unemployed at the beginning of the study and had valid information on benefit payments. Actual estimation is done only for those who were unemployed for at least four weeks. As only the first spell of registered unemployment is included in the estimation, there are as many spells as individuals in the sample.

Initial model development prior to the inclusion of unobserved heterogeneity is similar to MEYER (1990), but NARENDRANATHAN/STEWART (1993B) utilise the connection to binary response models as suggested by equation (13) above and also estimate probit and logit formulations, though they are not directly implied by the Proportional Hazards model. They find that both the probit and the logit-model dominate the extreme value-formulation in likelihood terms. The results suggest no benefit effect past the twelfth week of a spell and a declining effect within the first twelve weeks.

Since the estimated elasticities from the logit model are found to lie between the values obtained from the other two formulations, the further analysis, now including unobserved heterogeneity, is done on the basis of the logit specification. Following HECKMAN/SINGER (1982, 1984A, 1984B, 1985, 1986), a set of two mass points is taken as a discrete approximation of the distribution of the heterogeneity component. Alternatively, the component is assumed to be normally distributed, which is justified by the Central Limit Theorem. However, results for both mixture specifications are very similar. Now the benefit effect is significant up to the twentieth week with a steady decline in the effect up to this point. Thus, omitted heterogeneity seems to intensify the negative duration dependence in the unemployment income elasticity.

An interesting econometric issue discussed by NARENDRANATHAN/STEWART (1993B) is the assumption of independence between the covariates and the error term, if variables capturing the individual' s previous labour market experience are included in the specification. There might be a correlation between those variables and the heterogeneity component. On the other hand, excluding those variables to avoid an endogeneity bias can also cause a serious misspecification as the previous labour market experience clearly influences the transition probability.

In the context of modelling the probability of leaving unemployment, transition into employment is not the only way of leaving unemployment. Individuals might simply withdraw completely from the labour market, retire because of age, illness or disability, etc. as has already been discussed in section II.3 above. **NARENDRANATHAN/STEWART (1993A)** extend their model just outlined to the case of competing risks using the same data and distinguishing between an "exit hazard" and a "return-to-employment hazard".

Estimation in a dependent competing risks environment along the lines of II.4.4 is done by **HAN/HAUSMAN (1990)** following KATZ'S (1986) paper based on the Panel Study of Income Dynamics (PSID), who had studied the determinants of unemployment duration and divided the hazards into either new jobs or recalls. Unlike HAN/HAUSMAN (1990), KATZ (1986) had assumed independence of the risks and specified the baseline hazard as being of a Weibull type. Like MEYER (1987, 1990) or NARENDRANATHAN/STEWART (1993A,B), HAN/HAUSMAN (1990) instead conduct nonparametric estimation of the baseline hazard. Their model for both hazards is specified as a two-equation-GAFT-model

$$\begin{aligned} t_1^* &= -\ln\int_0^{t_1} \lambda_0^1(s)\,ds = x'\beta_1 + \varepsilon_1 \\ t_2^* &= -\ln\int_0^{t_2} \lambda_0^2(s)\,ds = x'\beta_2 + \varepsilon_2 \end{aligned} \qquad (21).$$

To allow for stochastic dependence between the two hazards HAN/HAUSMAN (1990) assume the joint distribution function of $\varepsilon_1$, $\varepsilon_2$ to be joint standard normal, which leads to a bivariate ordered probit model. As the authors admit, this is only an approximation to the proportional hazards specification, but has the advantage of permitting unrestricted correlation between the disturbances.

The results of HAN/HAUSMAN (1990) show – like KATZ (1986) – significantly different hazards for the two types of risks, but reject his baseline hazard specification. His finding of strong positive duration dependence in the case of the new job hazard appears to be the result of the Weibull specification rather than actual individual behaviour. His assumption of stochastic independence, however, is not rejected at usual significance levels.

Another empirical application where dependence play an important role is **BELZIL (1995)**, who analyses the impact of unemployment insurance on the likelihood of *re-entering* unemployment: On one hand, a more generous unemployment compensation tends to increase reservation wages, thereby, perhaps, significantly improving subsequent job matching and employment duration. On the other hand, longer unemployment durations, implied by higher benefits, might have adverse effects on reemployment wages and durations. Clearly, in this framework, because reemployment duration depends inter alia on the completed duration of unemployment, in the presence of unobserved heterogeneity, the unemployment and reemployment equations must be specified as a simultaneous system. The model ultimately estimated by BELZIL (1995) using data from the Longitudinal Labor Force File of Employment and Immigration Canada 1972-1984, however, is a simple recursive and linear two-equation-model with the logarithm of duration spent in unemployment (employment) $t_u$ ($t_e$) as the dependent variable

$$\begin{aligned} \log t_u &= x_u'\beta_u + \varepsilon_u \\ \log t_e &= x_e'\beta_e + \varepsilon_e \end{aligned} \qquad (22),$$

where $x_u$ contains the logarithm of the level and of the potential duration of unemployment benefits when unemployment began and $x_e$ includes $\log t_u$ as explanatory variables. Again, the error terms are assumed to be jointly normal. Formulation (21) implies a particular parametric specification of the hazard function: It corresponds to an exponential model for the hazard since the baseline hazard is equal to unity. BELZIL (1995) presents estimates both for the complete sample and for subsamples based on a distinction between recalls and new jobs. These competing risks are obviously treated as independent from each other since both subsamples are analysed "separately" (BELZIL (1995), p.121). BELZIL'S (1995) results sustain the hypothesis that unemployment benefits increase unemployment durations and that unemployment duration decrease reemployment durations. Also, his findings indicate "that the unemployment durations of those who have obtained a new job are more sensitive" to unemployment insurance parameters "and that their reemployment durations are more sensitive to completed unemployment duration" (BELZIL (1995), p.124).

**HUJER/SCHNEIDER (1996)** analyse inter alia determinants of the transitions from unemployment to either employment or non-employment for women in the Socio-Economic Panel for West

Germany 1983-1992. They too assume independence between both transitions. Using a discrete hazard model with unobserved heterogeneity following a gamma distribution, they find that young, high qualified and experienced unemployed have the best chances of obtaining a new job. An interesting result is also that the transition into non-employment seems to be influenced by moral-hazard considerations.

The focus of **STEINER (1994)** is on the relative importance of *state dependence* versus *sorting effects* for explaining the duration and persistence of unemployment. Under the former theory, employment probabilities might deteriorate with the duration of the ongoing unemployment spell because of, for example, loss of human capital (*negative duration dependence*). Instead of or in addition to the present unemployment spell, previous unemployment experience could have the same effect (*lagged duration*, *occurrence dependence*). Sorting effects can explain rising long-term unemployment even if individual re-employment probabilities do not deteriorate with unemployment duration: The proportion of people with low re-employment probabilities might increase during and after a recession thereby leading to a rising share of long-term unemployment (STEINER (1994), p.2).

To answer his question, STEINER (1994) analyses the transitions from unemployment into employment and – for females only – non-participation, which he assumes to be independent from each other, on the basis of waves 1-9 of the Socio-Economic Panel for West Germany using a multinomial logit model with a set of three mass points as a discrete approximation of the heterogeneity distribution. His estimates indicate that there is no decline in individual re-employment probabilities for males and the great majority of females and instead point to sorting effects as the cause for incereasing unemployment persistence.

In Eastern Germany *Arbeitsbeschaffungsmaßnahmen* (ABM; employment schemes) are an important policy instrument to prevent unemployment in a narrow sense. The question, whether participation in those employment schemes has a positive reemployment effect, is examined by **STEINER/KRAUS (1995)** using data for 1990-1992 from the *Arbeitsmarktmonitor* (labour market monitor), a panel study undertaken by the *Bundesanstalt für Arbeit* (German Federal Bureau of Labor) in East Germany between 1990 and 1994. Since data in the Arbeitsmarktmonitor for the duration/transition are only measured on a monthly basis, STEINER/KRAUS (1995) define their model in discrete time. Their econometric model is basically identical to that used by STEINER (1994): They specify separate multinomial logit-models for the transition from unemployment into employment, ABM and non-employment on one hand and for the transition from ABM into employment and non-employment (including unemployment) on the other. By comparing the transition probabilities into employment from unemployment with those from ABM, STEINER/KRAUS (1995) arrive at the result that for women the participation effect is negative whereas for men there is a short-run positive effect after completion of the scheme.

STEINER/KRAUS (1995) think that by incorporating unobserved heterogeneity (of Heckman/Singer type) they can overcome the problem of sample selection. Though they do not have the problem of a possible correlation between the heterogeneity component and the participation status of the individual because participation status is not an exogenous variable in their model, their comparison between transition probabilities, however, is still biased by sample selection (see HUJER/MAURER/WELLNER (1996) for a more complete discussion). The fact that they also estimate the transition into ABM does not help in that respect as it does not influence the estimates for the transition from ABM into employment.

**HAM/LALONDE (1996)** have the advantage of being able to rely on a social experiment (National Supported Work Demonstration for women in seven cities in the US 1976/1977) in order to analyse the effects of a training program on the duration of participants' subsequent employment and unemployment spells. Thus, they avoid the problem of a correlation between training status and heterogeneity component and do not need to simultaneously model the selection process into training in addition to "the process that generates the outcomes of interest" (HAM/LALONDE (1996), p.177). HAM/LALONDE (1996) direct their attention to transitions between employment and unemployment for the control group and between training, employment and unemployment for the experimental group. In their more advanced specifications they allow for correlation between the cause-specific transition rates, which are basically of simple logit type. Heterogeneity is incorporated following HECKMAN/SINGER (1982, 1984A, 1984B, 1985, 1986). They find "that NSW raised employment rates because it helped women who found jobs remain employed longer than they would have otherwise" (HAM/LALONDE (1996), p.199). However, the program had no significant effect on the duration of unemployment spells.

# IV. Conclusion

This paper has surveyed methods for the evaluation of grouped transition data as well as recent empirical applications of these methods – though this separation is rather arbitrary, because, as it is often the case in econometrics, the innovation of theoretical methods usually coincides with their empirical application. Of course, time is a continuous phenomenon and therefore time-continuous hazard models are the basis for models for grouped transition data, but the latter models take explicitly into account the fact that economic processes and decisions generally are only observed at discrete points in time, thereby avoiding the possibility of biased estimates resulting from too many ties.

Perhaps, the most intriguing aspects of models for grouped transition data are

(1) the shape of the discrete hazard rate developed from the Proportional Hazards model without any distributional assumptions, which is identical to the extreme value distribution;

(2) the similarity to binary/ordered response models.

As we have seen, both aspects often lead to the specification and estimation of these related models in empirical application. Results like those of NARENDRANATHAN/STEWART (1993B) seem to indicate that the error thus made can be neglected. Binary or ordered response models have the advantage of being far more wide-spread and often of being easier to apply. The work of SUEYOSHI (1995), who compares the implications for the hazard behaviour of a logit, normal and extreme value specification for interval specific survivor functions, however, "suggests that some care should be taken to investigate the assumptions embodied in a particular specification of the conditional exit probabilities" (SUEYOSHI (1995), p.430).

Up to now, most empirical applications of hazard models for grouped (but also for continuous) data focus on unemployment benefit effects. Only a few tried to analyse the effects of training on employment histories using these methods. The examination of training effects is usually more complicated because of the possibility of sample selection bias, at least in a non-experimental framework. Experimental designs have other serious drawbacks. For instance, the effects one has arrived at by examining social experiments can not be easily transferred to the labour market or training in general, because social experiments are normally limited to specific training programs. Thus, our analysis of training effects using a panel study like the Socio-Economic Panel might provide further insight into the working of training if we manage to adequately account for sample selection effects.

# Literature

BELZIL, CHRISTIAN (1995): Unemployment Insurance and Unemployment Over Time: An Analysis with Event History Data, in: *Review of Economics and Statistics*, Vol.77, No.1, p.113-126.

BLOSSFELD, HANS-PETER/HAMERLE, ALFRED/MAYER, KARL ULRICH (1986): Ereignisanalyse, Frankfurt and New York.

COX, D.R. (1972): Regression Models and Life-Tables (with discussion), in: *Journal of the Royal Statistical Soceity, Series B*, Vol.34, No.2, p.187-220.

COX, D.R./OAKES, D. (1984): Analysis of Survival Data, London, New York.

ELBERS, CHRIS/RIDDER, GEERT (1982): True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model, in: *Review of Economic Studies*, Vol.49, No.3, p.403-409.

GRITZ, MARK R. (1993): The Impact of Training on the Frequency and Duration of Employment, in: *Journal of Econometrics*, Vol.57, p.21-51.

HAM, JOHN C./LALONDE, ROBERT J. (1996): The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training, in: *Econometrica*, Vol.64, No.1, p.175-205.

HAMERLE, ALFRED/TUTZ, GERHARD (1989): Diskrete Modelle zur Analyse von Verweildauern und Lebenszeiten, Frankfurt and New York.

HAN, AARON/HAUSMAN, JERRY A. (1990): Flexible Parametric Estimation of Duration and Competing Risk Models, in: *Journal of Applied Econometrics*, Vol.5, p.1-28.

HECKMAN, JAMES J./SINGER, BURTON (1982): The Identification Problem in Econometric Models for Duration Data, in: Werner Hildenbrand (ed.), Advances in Econometrics, Cambridge, p.39-77.

HECKMAN, JAMES J./SINGER, BURTON (1984A): A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data, in: *Econometrica*, Vol.52, No.2, p.271-320.

HECKMAN, JAMES J./SINGER, BURTON. (1984B): Econometric Duration Analysis, in: *Journal of Econometrics*, Vol.24, p.63-132.

HECKMAN, JAMES J./SINGER, BURTON (1985): Social Science Duration Analysis, in: James J. Heckman and Burton Singer (eds.), Longitudinal analysis of labor market data, Cambridge, p.39-110.

HECKMAN, JAMES J./SINGER, BURTON (1986): Econometric Analysis of Longitudinal Data, in: Zvi Griliches and Michael D. Intriligator (eds.), Handbook of Econometrics, Vol. III, Amsterdam, p.1689-1763.

HUJER, REINHARD/SCHNEIDER, HILMAR (1989): The Analysis of Labor Market Mobility Using Panel Data, in: *European Economic Review*, Vol.33, p.530-536.

HUJER, REINHARD/SCHNEIDER, HILMAR (1996): Institutionelle und strukturelle Determinanten der Arbeitslosigkeit in Westdeutschland: Eine mikroökonometrische Analyse mit Paneldaten, in: Arbeitslosigkeit und Möglichkeiten ihrer Überwindung, Schriften des Wirtschaftswissenschaftlichen Seminars Ottobeuren, Vol.25, ed. by Bernhard Gahlen, Helmut Hesse and Hans Jürgen Ramser, Tübingen, p.53-76.

HUJER, REINHARD/MAURER, KAI-OLIVER/WELLNER. MARC (1996): The Impact of Training on Employment: A Survey of Microeconometric Studies, Frankfurter Volkswirtschaftliche Diskussionsbeiträge No.69, Department of Economics, Johann Wolfgang Goethe-University, Frankfurt/Main.

JOHNSON, NORMAN L./KOTZ, SAMUEL (1970): Continuous Univariate Distributions - 1, New York.

KALBFLEISCH, J./PRENTICE, R. (1980): The Statistical Analysis of Failure Time Data, New York.

KATZ, LAWRENCE F. (1986): Layoffs, Recall and the Duration of Unemployment, NBER Working Paper No.1825.

KIEFER, NICHOLAS M. (1988A): Economic Duration Data and Hazard Functions, in: *Journal of Economic Literature*, Vol.26, p.646-679.

KIEFER, NICHOLAS (1988B): Analysis of Grouped Duration Data, in: N.U. Prabhu (ed.), Statistical Inference from Stochastic Processes, Contemporary Mathematics, Vol.80, Providence, p.107-137.

KIEFER, NICHOLAS.M. (1990): Econometric Methods for Grouped Duration Data, in: Joop Hartog, Geert Ridder and Jules Theeuwes (eds.), Panel Data and Labor Market Studies, Amsterdam, p.97-117.

LANCASTER, TONY (1979): Econometric Methods for the Duration of Unemployment, in: *Econometrica*, Vol.47, No.4, p.939-956.

LANCASTER, TONY (1990): The Econometric Analysis of Transition Data, Cambridge.

MADDALA, G.S. (1983): Limited-dependent and Qualitative Variables in Econometrics, Cambridge.

MEYER, BRUCE D. (1987): Hazard and Markov Chain Models with Applications to Labor Economics, M.I.T. Ph.D. Thesis.

MEYER, BRUCE D. (1990): Unemployment Insurance and Unemployment Spells, in: *Econometrica*, Vol.58, No.4, p.757-782.

NARENDRANATHAN, WIJI/STEWART, MARK B. (1993A): Modelling the Probability of Leaving Unemployment: Competing Risks Models with Flexible Base-line Hazards, in: *Applied Statistics*, Vol.42, p.63-83.

NARENDRANATHAN, WIJI/STEWART, MARK B. (1993B): How Does the Benefit Effect Vary as Unemployment Spells Lengthen?, in: *Journal of Applied Econometrics,* Vol.8, p.361-381.

RIDDER, GEERT (1990): The Non-Parametric Identification of Generalized Accelerated Failure-Time Models, in: *The Review of Economic Studies*, Vol.57, No.2, p.165-182.

STEINER, VIKTOR (1994): Labour Market Transitions and the Persistence of Unemployment – West Germany 1983 -1992, Discussion Paper No.94-20, Zentrum für Europäische Wirtschaftsforschung, Mannheim.

STEINER, VIKTOR/KRAUS, FLORIAN (1995): Haben Teilnehmer an Arbeitsbeschaffungsmaßnahmen in Ostdeutschland bessere Wiederbeschäftigungschancen als Arbeitslose?, in: Viktor Steiner and Lutz Bellmann (eds.), Mikroökonomik des Arbeitsmarktes, Beiträge aus der Arbeitsmarkt- und Berufsforschung, Vol.192, Nürnberg, p.387-423.

SUEYOSHI, GLENN T. (1994): Semiparametric Estimation of Generalized Accelerated Failure Time Models with Grouped Data, Discussion Paper 94-10, Department of Economics, University of California, San Diego.

SUEYOSHI, GLENN T. (1995): A Class of Binary Response Models for Grouped Duration Data, in: *Journal of Applied Econometrics*, Vol.10, p.411-431.

TRUSSELL, J./RICHARDS, T. (1985): Correcting for Unmeasured Heterogenity in Hazard Models Using the Heckman-Singer Procedure, in: Nancy Brandon Tuma (ed.), Sociological Methodology, San Francisco, London, Washington, p.242-276.

TUMA, NANCY BRANDON/HANNAN, MICHAEL T. (1984): Social Dynamics - Models and Methods, Orlando.