Brecht, Beatrix; Brecht, Leo

Working Paper

Time discrete nonparametric survival analysis using panel data

Diskussionsbeiträge - Serie II, No. 200

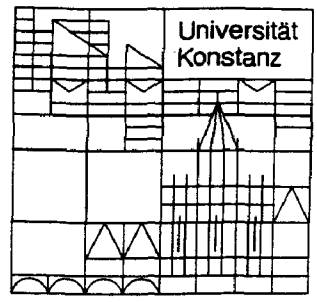
Provided in Cooperation with:
Department of Economics, University of Konstanz

Suggested Citation: Brecht, Beatrix; Brecht, Leo (1993) : Time discrete nonparametric survival analysis using panel data, Diskussionsbeiträge - Serie II, No. 200, Universität Konstanz, Sonderforschungsbereich 178 - Internationalisierung der Wirtschaft, Konstanz

This Version is available at:
https://hdl.handle.net/10419/101509

Sonderforschungsbereich 178
„Internationalisierung der Wirtschaft“

Diskussionsbeiträge



Juristische
Fakultät

Fakultät für Wirtschafts-
wissenschaften und Statistik

Beatrix Brecht
Leo Brecht

Time-discrete Nonparametric Survival Analysis Using Panel Data

---

Postfach 5560
D-7750 Konstanz

Serie II — Nr. 200
März 1993

TIME-DISCRETE NONPARAMETRIC SURVIVAL ANALYSIS USING PANEL DATA

Beatrix Brecht

Leo Brecht

Serie II - Nr. 200

März 1993

TIME-DISCRETE NONPARAMETRIC SURVIVAL ANALYSIS USING PANEL DATA

Beatrix Brecht and Leo Brecht
SFB 178, University of Konstanz
and
Institute for Information Management,
University of St. Gallen

Abstract:
In this contribution a nonparametric estimator for the hazard function will be presented for time-discrete survival analysis. The estimator is derived from a likelihood function based upon time-discrete counting processes. With martingale techniques asymptotic properties of the estimator of the cumulative hazard function are shown. Since we consider a nonparametric approach we can only estimate remigrant behaviour referring to the length of stay, causes which are due to the differences are not part of this investigation. The estimations are carried out with a module of PRODISA, a program package developed for the analysis of time-discrete duration and panel data for the nonparametric and (semi)parametric case. For analysing the remigrant behaviour of different foreign nations (Italy, Yugoslavia, Greece, Spain and Turkey) the Socio-Economic Panel (SOEP) is used as a data basis.

1 Foundations

In many cases, no exact points of time can be given at which events or transitions occur, only time intervals. If one nevertheless uses a continuous-time model, this causes a great number of similar observation values - so called "ties" - to appear, and one generally obtains useless parameter estimations. Since the theoretical foundations, such as the derivation of asymptotic characteristics, proceed from the assumption that no ties are present, the theoretical foundation of this statistical method is invalid if a continuous-time model is still taken as a basis. For the modelling of discrete-time raised event times,

the time axis is divided into T_0 intervals $(a_0, a_1]$, $(a_1, a_2], \cdots, (a_{T_0-1}, \infty)$. The intervals themselves are numbered from $t = 1, 2, \cdots, T_0$. One can therefore observe a probability model with a positive random variable T, which can take on integer values from the set $\{1, \cdots, T_0\}$. $\{T = t\}$ means that a transition has taken place in the interval $t = (a_{t-1}, a_t]$. The individuals that are at risk at the beginning of the interval t remain at risk during the entire interval, and events that occur during the interval t are interpreted as if they had occurred at point a_t.

Let $T_1, \ldots, T_n$ be independently and identically distributed random variables with a given discrete probability density.
The hazard function for the discrete case can be defined as follows:

$$\lambda(t) = P(T = t \mid T \geq t) \quad \text{for } t = 1, \ldots, T_0 \tag{1.1}$$

$T = t$ indicates that a transition has occurred in this interval. In (1.1), the conditional probability is rendered so that an observed individual undergoes a transition in the time interval t, given that the individual reached the beginning of the time interval.

To simplify the notation, the one-state case will be considered, which can be easily generalized to k different events.

The case of censored event times often occurs, i.e. no event occurs during the period of observation, or the individual is no longer available for the study. In that case, the tupel $(T_1, \delta_1), \cdots, (T_n, \delta_n)$ is given with $T_i = min(I_i, C_i)$ and with the censoring indicator δ_i, in which $\delta_i = 1$ or 0, depending upon whether I_i is observed or not, and I_i and C_i are discrete positive random variables.
With the definition of the indicator function given in

$$\begin{aligned} Y_i(t) &= I\{T_i \geq t\}, \quad i = 1, \cdots, n, \\ N_i(t) &= I\{T_i \leq t; \delta_i = 1\}, \quad i = 1, \cdots, n, \end{aligned}$$

it is possible to indicate a closed term for the likelihood function of a discrete sojourn model. The indicator function $Y_i(t)$ is to be interpreted here in such a manner that it retains the value one as long as no event has occurred for the i-th individual. It holds true for the function $N_i(t)$ that the value one is only reached when an event has occurred. Furthermore, in order to formulate the likelihood function, one requires the concept of the risk set, which is given by

$$R(t) = \{i : Y_i(t) = 1\}. \tag{1.2}$$

Furthermore,

$$N(t) = \sum_{i=1}^{n} N_i(t), \quad Y(t) = \sum_{i=1}^{n} Y_i(t), \quad t = 1, \cdots, T_0, \tag{1.3}$$

holds, and $\Delta N_i(t)$ is defined as

$$\Delta N_i(t) = N_i(t) - N_i(t-1), \; N_i(0) = 0. \tag{1.4}$$

$\Delta N_i(t)$ should indicate whether a transition has occurred for the i-th individual in the t-th interval ($\Delta N_i(t) = 1$), or not ($\Delta N_i(t) = 0$).
Based on various assumptions, *Arjas/Haara (1987)* then define a likelihood function according to

$$L = \prod_{t \le T_0} \prod_{i \in R(t)} P(\Delta N_i(t) = \Delta n_i(t) \mid \mathcal{G}_{t-1}). \tag{1.5}$$

There, the σ-algebra $\mathcal{G}_{t-1}$ contains the previous history of the process up until the interval $t-1$.

The function $N(t)$, which has been introduced, is the counting process that indicates the number of events that have occurred in the interval $[0, t)$, whereby $t = 1, \ldots T_0$ once again holds. The number of individuals still exposed to risk is given by the represented function $Y(t)$. The following relations can therefore be ascertained for the time-discrete nonparametric model:
The cumulative hazard function is

$$\lambda(t) = \sum_{k=1}^{t} \lambda(k) \quad \text{for} \quad t \le T_0. \tag{1.6}$$

The survivor function correspondingly results to

$$S(t) = \prod_{k=1}^{t-1} (1 - \lambda(k)). \tag{1.7}$$

By carrying over the likelihood function to the nonparametric time-discrete method, a nonparametric estimator can be derived for the hazard function. For the likelihood

$$L = \prod_{t=1}^{T_0} \lambda(t)^{\Delta N(t)} (1 - \lambda(t))^{Y(t) - \Delta N(t)}, \tag{1.8}$$

with

$$\Delta N(t) = \sum_{i=1}^{n} \Delta N_i(t),$$

the hazard function estimator

$$\hat{\lambda}(t) = \begin{cases} \dfrac{\Delta N(t)}{Y(t)} & \text{for } Y(t) > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{1.9}$$

results whereby for $\Delta N(t)$,

$$\Delta N(t) = N(t) - N(t-1), \quad N(0) = 0 \tag{1.10}$$

holds.

If one uses an additional indicator function $K(t)$, that retains the value one in the case of $Y(t) > 0$ and otherwise takes on the value zero, i.e. $K(t) = I\{Y(t) > 0\}$, the notation is shortened to

$$\hat{\lambda}(t) = \frac{\Delta N(t)}{Y(t)} K(t). \tag{1.11}$$

The hazard function estimator thus obtained can be used in (1.6) so that the estimator

$$\hat{\Lambda}(t) = \sum_{k=1}^{t} \frac{\Delta N(k)}{Y(k)} K(t) \tag{1.12}$$

can be given for the cumulative hazard function. Correspondingly, one obtains

$$\hat{S}(t) = \prod_{k=1}^{t-1} \left(1 - \frac{\Delta N(k)}{Y(k)}\right) K(t). \tag{1.13}$$

as an estimator for the survivor function from (1.7). If one uses the model of an alternative cumulative hazard function

$$\tilde{\Lambda}(t) = \sum_{k=1}^{t} K(k)\lambda(k), \tag{1.14}$$

suggested by Anderson/Borgan (1985) for the time continuous case, which differs from the cumulative hazard function from (1.6) by the additionally considered indicator function $K(t)$, then asymptotic statements about the properties of the cumulative hazard function estimator can be made using martingale techniques. In martingale theory the conditional variance of a martingale M is given by the so called predictable variance process given by

$$\langle M \rangle(t) = Var(M(t) \mid G_{t-1}) \tag{1.15}$$

and the predictable covariance process of two martingales M_1,M_2

$$\langle M_1, M_2 \rangle(t) = Cov(M_1(t), M_2(t) \mid G_{t-1}). \tag{1.16}$$

Assumption 1: *(Conditional Independence of Transitions)*
For each $t \geq 1$, the random variables $\{\Delta N_i(t); i \geq 1\}$, $i \in \mathcal{R}(t)$, are independent under the condition $\mathcal{G}_{t-1}$.

Lemma 1.1 *Let $\Delta M_i(t) := \Delta N_i(t) - Y_i(t)\lambda(t)$ then we get: $M_i(t)$ is a martingale with respect to G_{t-1}.*

Proof

$$\begin{aligned} E(\Delta M_i(t) \mid \mathcal{G}_{t-1}) &= P(\Delta N_i(t) = 1 \mid \mathcal{G}_{t-1}) - Y_i(t)\lambda(t) \\ &= Y_i(t)\lambda(t) - Y_i(t)\lambda(t) = 0 \end{aligned} \tag{1.17}$$

q.e.d.

Assumption 1 simplifies the handling of ties.

If one substitutes

$$\begin{aligned} \Delta\langle M_i\rangle(t) &= Var(\Delta M_i(t) \mid \mathcal{G}_{t-1}), \\ \Delta\langle M_i, M_j\rangle(t) &= Cov(\Delta M_i(t), \Delta M_j(t) \mid \mathcal{G}_{t-1}), \end{aligned}$$

one obtains

Lemma 1.2
With assumption 1 it results:
(i)

$$\Delta\langle M_i\rangle(t) = Y_i(t)\lambda(t)(1 - \lambda(t)) \tag{1.18}$$

(ii)

$$\Delta\langle M_i, M_j\rangle(t) = 0, \quad i \neq j, \tag{1.19}$$

i.e. the martingales M_i and M_j are orthogonal for $i \neq j$.

Proof

(i) Follows by using

$$P(\Delta N_i(t) = 1 \mid \mathcal{G}_{t-1}) = Y_i(t)\lambda(t).$$

(ii)

$$
\begin{aligned}
&Cov(\Delta N_i(t) - Y_i(t)\lambda(t), \Delta N_j(t) - Y_j(t)\lambda(t)|\mathcal{G}_{t-1}) \\
&= Cov(\Delta N_i(t), \Delta N_j(t)|\mathcal{G}_{t-1}) \\
&= E(\Delta N_i(t) \cdot \Delta N_j(t)|\mathcal{G}_{t-1}) - E(\Delta N_i(t)|\mathcal{G}_{t-1}) \cdot E(\Delta N_j(t)|\mathcal{G}_{t-1}) \\
&= 0.
\end{aligned}
$$

There the last equation follows from assumption 1. q.e.d.

One obtains

Lemma 1.3

Under assumption 1 it holds:

(i) $\hat{\Lambda}(t) - \tilde{\Lambda}(t)$ *is a local square integrable martingale referring to* $\mathcal{G}_{t-1}$ *with the expected value* 0.

(ii) *For the predictable variance process*

$$\langle \hat{\Lambda} - \tilde{\Lambda},\ \hat{\Lambda} - \tilde{\Lambda} \rangle(t) = \sum_{\nu=1}^{t} \frac{K(k)}{Y(k)} \lambda(k)(1 - \lambda(k))$$ *holds.*

(iii) *The estimator of the mean square error function*

$MSE(t) = E(\hat{\Lambda}(t) - \tilde{\Lambda}(t))^2$ *defined to*

$$\widehat{MSE}(t) = \sum_{k=1}^{t} \frac{K(k)}{(Y(k))^2} \Delta N(k)$$ *is unbiased.*

Proof

(i) For $\hat{\Lambda}(t) - \tilde{\Lambda}(t)$ one obtains

$$\hat{\Lambda}(t) - \tilde{\Lambda}(t) = \sum_{\nu=1}^{t} \frac{K(\nu)}{Y(\nu)} \Delta M(\nu)$$

with $\Delta M(\nu) = \Delta N(\nu) - Y(\nu)\lambda(\nu)$; therefore we get

$$E(\hat{\Lambda}(t) - \tilde{\Lambda}(t) \mid \mathcal{G}_{t-1}) = \sum_{\nu=1}^{t} \frac{K(\nu)}{Y(\nu)} E(\Delta M(\nu) \mid \mathcal{G}_{t-1}) = 0.$$

The first equation follows from the predictability of $K(\nu)/Y(\nu)$, the second equation from the martingale property of $\Delta M(\nu) = \sum_{i=1}^{n} \Delta M_i(\nu)$, and with lemma 1.1 the statement for the time-discrete case follows.

(ii) For the variance process it holds $\langle \hat{\Lambda} - \tilde{\Lambda},\ \hat{\Lambda} - \tilde{\Lambda} \rangle(t)$

$$\begin{aligned} \langle \hat{\Lambda} - \tilde{\Lambda},\ \hat{\Lambda} - \tilde{\Lambda} \rangle(t) &= \sum_{\nu=1}^{t} \frac{K(\nu)}{(Y(\nu))^2} \Delta\langle M \rangle(\nu) \\ &= \sum_{\nu=1}^{t} \frac{K(\nu)}{Y(\nu)} \lambda(\nu)(1 - \lambda(\nu)). \end{aligned}$$

(iii) For the mean square error function it follows

$$MSE(t) = E(\hat{\Lambda}(t) - \tilde{\Lambda}(t))^2 = E(\langle \hat{\Lambda} - \tilde{\Lambda} \rangle(t))$$

and therefore as an estimator for $MSE(t)$ one obtains $\sum_{\nu=1}^{t} \frac{K(\nu)}{(Y(\nu))^2} \Delta N(\nu)$. By taking the expectation of

$$\hat{\Lambda}(t) - MSE(t) = \sum_{\nu=1}^{t} \frac{K(\nu)}{(Y(\nu))^2} \Delta M(\nu)$$

q.e.d.

we get the unbiasness.

In order to study the asymptotic properties of these time-discrete estimators, let us consider a sequence of counting processes indicated with $n = 1, 2, \ldots$, which all satisfy the time-discrete hazard model.

Lemma 1.4

Under the assumption

$$\sum_{\nu=1}^{t} \frac{K^{(n)}(\nu)}{Y^{(n)}(\nu)} (1 - \lambda(\nu))\lambda(\nu) \xrightarrow{p} 0, \quad n \longrightarrow \infty, \text{ for all } \quad t \in \{1, \ldots, T_0\},$$

it follows

$$\max_{\nu \in \{1,\ldots,T_0\}} | \hat{\Lambda}(\nu) - \tilde{\Lambda}(\nu) | \xrightarrow{p} 0, \quad n \longrightarrow \infty.$$

Proof

As $\hat{\Lambda}(t) - \tilde{\Lambda}(t) = \sum_{\nu=1}^{t} \frac{K^{(n)}(\nu)}{Y^{(n)}(\nu)} \Delta M(\nu)$ is a local square integrable martingale it follows with Lenglart's inequality (see Appendix)

$$P(\max_{\nu \in \{1,\ldots,T_0\}} | \hat{\Lambda}(\nu) - \tilde{\Lambda}(\nu) | > \delta) \leq \frac{\delta}{\eta^2} + P(\langle \hat{\Lambda} - \tilde{\Lambda} \rangle(\nu) > \eta)$$

and with the assumption it can be shown

$$(\hat{\Lambda} - \tilde{\Lambda})(t) = \sum_{\nu=1}^{t} \frac{K^{(n)}(\nu)}{Y^{(n)}(\nu)}(1 - \lambda(\nu))\lambda(\nu) \xrightarrow{p} 0, \; n \longrightarrow \infty.$$

q.e.d.

To prove the asymptotic properties of $\hat{\Lambda}(t)$, one needs the following regularity conditions:

$$\text{(D1)} \quad \sum_{k=1}^{t} n\frac{K^{(n)}(k)}{Y^{(n)}(k)}\lambda(k)(1-\lambda(k)) \xrightarrow{p} \sum_{k=1}^{t} h(k), \; n \to \infty \; ,$$

$$\text{(D2)} \quad \sum_{k=1}^{t} n\frac{K^{(n)}(k)}{Y^{(n)}(k)}\lambda(k)(1-\lambda(k))I\left\{ \mid \sqrt{n}\frac{K^{(n)}(k)}{Y^{(n)}(k)} \mid > \epsilon \right\} \xrightarrow{p} 0, \; n \to \infty,$$

for all t. With these conditions one formulates the theorem

Theorem 1.1

Under the regularity conditions (D1) and (D2),

$$\sqrt{n}\left(\hat{\Lambda}^{(n)}(t) - \tilde{\Lambda}^{(n)}(t)\right) \xrightarrow{d} W(t),$$

holds; in addition, W represents an independent Gaussian martingale with $W(0) = 0$ and

$$Cov(W(t), W(s)) = \sum_{k=1}^{t \wedge s} h(k).$$

Proof

One defines

$$V^{(n)}(t) = \sqrt{n}(\hat{\Lambda}^{(n)}(t) - \tilde{\Lambda}^{(n)}(t)) = \sqrt{n}\sum_{\nu=1}^{t} \frac{K^{(n)}(\nu)}{Y^{(n)}(\nu)}\Delta M(\nu),$$

$$\bar{I}^{n\epsilon}(t) = I\left\{ \mid \sqrt{n}\frac{K^{(n)}(t)}{Y^{(n)}(t)} \mid > \epsilon \right\},$$

$$\bar{V}^{(n)\epsilon}(t) = \sqrt{n}\sum_{\nu=1}^{t} \frac{K^{(n)}(\nu)}{Y^{(n)}(\nu)}\bar{I}^{\epsilon}(\nu)\Delta M(\nu).$$

Then $\bar{V}^{(n)\epsilon}(t)$ is a jump part of an ϵ-decomposition of the martingale $V^{(n)}(t)$ (see Gill (1980)) and one obtains

$$V^{(n)}(t) = \bar{V}^{(n)\epsilon}(t) + (V^{(n)}(t) - \bar{V}^{(n)\epsilon}(t)).$$

Therefore it has to be shown

$$\begin{aligned} &\text{(i)} \quad \langle V^{(n)}, V^{(n)} \rangle(t) \quad \xrightarrow{p} \quad \textstyle\sum_{\nu=1}^{t} h(\nu), \\ &\text{(ii)} \quad \langle \bar{V}^{(n)\epsilon}, \bar{V}^{(n)\epsilon} \rangle(t) \quad \xrightarrow{p} \quad 0, \quad n \longrightarrow \infty. \end{aligned}$$

To (i):

With lemma 1.3 (ii) and referring to condition (D1) it holds true

$$\langle V^{(n)}, V^{(n)} \rangle(t) = n \sum_{\nu=1}^{t} \frac{K^{(n)}(\nu)}{Y^{(n)}(\nu)} \lambda(\nu)(1 - \lambda(\nu)) \quad \xrightarrow{p} \quad \sum_{\nu=1}^{t} h(\nu).$$

To (ii):

$$\langle \bar{V}^{(n)\epsilon}, \bar{V}^{(n)\epsilon} \rangle(t) = n \sum_{\nu=1}^{t} \frac{K^{(n)}(\nu)}{Y^{(n)}(\nu)} \bar{I}^{n\epsilon}(\nu) \lambda(\nu)(1 - \lambda(\nu)) \quad \xrightarrow{p} \quad 0,$$

referring to condition (D2).

The proof follows with the ϵ-decomposition theorem (see Appendix). q.e.d.

Sufficient conditions for (D1) and (D2) are:

($D1^*$) It holds:

$$\max_{k \in \{1,\ldots,T_0\}} \mid n \frac{K^{(n)}(k)}{Y^{(n)}(k)} \lambda(k)(1 - \lambda(k)) - h(k) \mid \xrightarrow{p} 0, \quad n \to \infty,$$

($D2^*$) It holds:

$$\max_{k \in \{1,\ldots,T_0\}} \mid \sqrt{n} \frac{K^{(n)}(k)}{Y^{(n)}(k)} \mid \xrightarrow{p} 0, \; n \to \infty.$$

Under the conditions (D1) and (D2) it follows furthermore:

$$n \frac{K^{(n)}(k)}{Y^{(n)}(k)} \Delta N(k) \xrightarrow{p} h(k). \tag{1.20}$$

An asymptotic variance for the cumulative hazard function is given through (1.20). In the continuous-time model tests were introduced by *Andersen et al. (1982)* with which hypotheses about the underlying hazard functions can be tested. By means of the time-discrete approach with the appertaining asymptotic theory, it now becomes possible to construct a corresponding testing theory for the discrete model as well.

Estimations have been made with the previously cited model approaches for the time-discrete nonparametric sojourn analysis. In this case the software package PRODISA (Program for Discrete Survival Analysis) was used, which will be introduced with its functions in the following.

2 Data Basis and PRODISA

A. The Data Basis

The German Socio-Economic Panel (SOEP) of the DIW (Deutsches Institut für Wirtschaftsforschung), Berlin, is a representative micro-longitudinal data collection for Germany. The random sample, first started 1984, furnishes about 1000 variables yearly from surveys in which about 12,000 people in about 6000 households are asked every year. These variables include the composition of a household, occupation, mobility, income pattern, living conditions, as well as information about education, health or value judgements and satisfaction. A SOEP-East-Study was started in the former GDR in 1990, where about 2000 households and 4500 people were used as a basis.

In this study, six panel waves (1984-1989) from the SOEP-West were used, in which approximately 1400 households reported having a foreign head of household.

In order to ascertain whether there are differences between nationalities, one must determine when a remigration to the country of origin occurs by examining the collection of data. The coding of the data into two categories – 1. foreigners who have remigrated to their native countries and 2. foreigners who are visiting their native countries for a longer period of time – and taking the year of immigration to the country allows the length of stay to be shown in annual intervals.
However, since the portion of remigrating guest workers is relatively small, a higher portion of censored data is present. In order to keep the estimation from being biased, a subsample for each foreign group was formed so that the respective data set contains about 20 per cent censored data (see table 1).
This restriction should be taken into consideration when interpreting the results; as a control, an estimation with the entire sample appears to be meaningful.

For the practical data processing of the SOEP with the relational data base system INGRES, cf. *Brecht, B. (1990).* The program PRODISA, developed to carry out the estimations, is described in greater detail in the following section.

B. The Program PRODISA

The program PRODISA was developed for the analysis of time-discrete duration and panel data for the nonparametric and (semi)parametric case.

Table 1: Subsample from the DIW Data Basis

Native country	Number of Guest workers Remigrated Back (uncensored)	Number of Guest workers Staying in Germany (censored)	Total
Spain	75	19	94
Italy	105	26	131
Yugoslavia	41	10	51
Greece	61	15	76
Turkey	134	34	168

PRODISA distinguishes three classes of models:

1. Semiparametric time-discrete models
2. Nonparametric models
3. Panel models.

In the semiparametric case one can choose different time-discrete models and estimate each model either with time independent, time dependent or time dependent with time lagged covariates. Testing is carried out with Wald tests and others.
In the nonparametric case one can estimate using kernel estimation methods or nonparametric time-discrete estimates, the one applied in this paper.
In the panel case estimation is carried out using the method of Bye/Riley (1989) adjusting the matrix of the covariances. Testing is implemented in the usual way and one of the modules allows testing of unobserved heterogeneity.

The special features of PRODISA are:

- It covers a wide area of time-discrete survival analysis
- It has a standard input structure
- It models the influence of covariates very flexible
- It has very fast procedures useful for very large data sets.

In contrast to traditional computer programs for discrete survival analysis, this one is characterized by the fact that time-dependent covariates can be included in the modelling.

The program package GLAMOUR (developed at the University of Regensburg) assumes in the case of a discrete duration time that time-independent covariates are present, i.e. the covariates will not change their value between the observed intervals. This assumption is an important restriction of the modelling, especially in the case of socio-economic characteristics, for example the covariate of an individual's state of health or employment participation.

3 Nonparametric Time-discrete Analysis of Remigrant Behavior

For the empirical analysis, the duration times have been divided into five intervals as represented in table 2.

Table 2: Interval Division of the Length of Stay of Foreigners in Germany

Interval Number	1	2	3	4	5
Length of Stay in Years	1 – 8	9 – 13	14 – 19	20 – 25	$\geq$ 26

The selection of the interval division was made by means of an exploratory data analysis. In table 3 the results for the estimated hazard functions and survival functions are presented separately for each nationality.

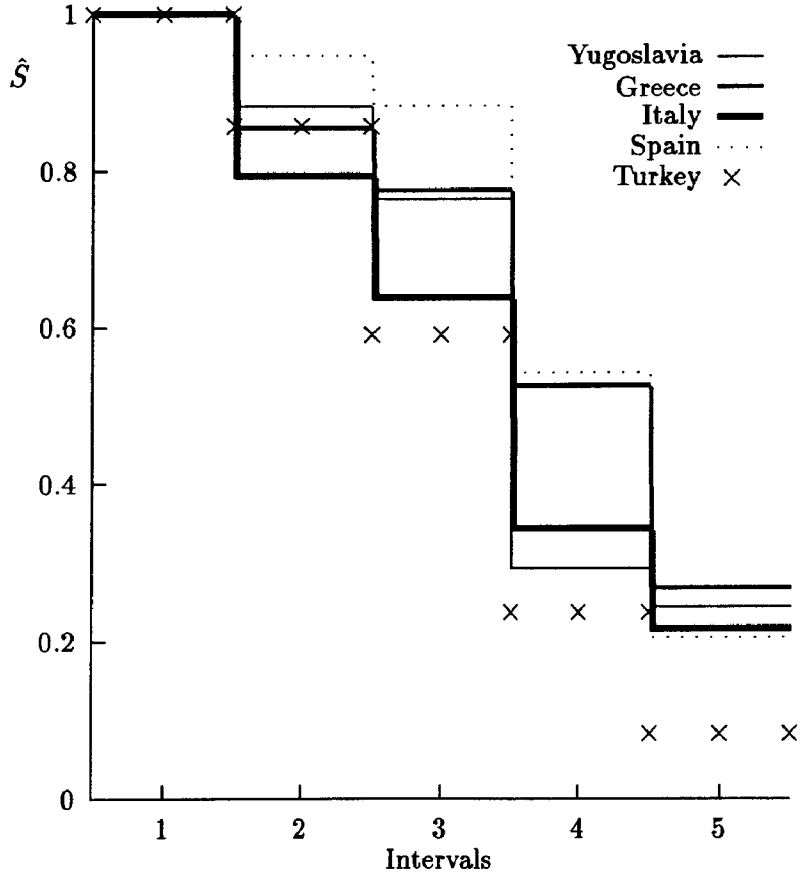
The remigration inclination seems to increase with the length of stay. Thus maximum values of the hazard function can be found in the last interval for Greece, Italy and Yugoslavia, i.e. with a duration of more than 25 years in Germany. In the case of Spain and Turkey, the desire to remigrate is the most pronounced in the fourth interval, however once again after a relatively long length of stay in the guest country. If one compares the survival functions (figure 1) of Spain with the one of Greece, one can see that at the beginning the Greek leave Germany earlier, but at the end more of them remain in Germany compared to Spanish guest workers. Different again are the values for the Turkish, the survival function decreases much more for them. An explanation for this that is often cited in literature is the attainment of a savings goal that should make a life as a self-employed worker in the native country possible, for example putting money away for a house or for the establishment of a business. Here, however, no covariates are taken into consideration; that is the task of other models.

Table 3: Estimation Results

Interval		Greece	Italy	Yugo-slavia	Spain	Turkey
1	Hazard	0.1447	0.2061	0.1177	0.532	0.1428
	Survival	1.000	1.000	1.000	1.000	1.000
2	Hazard	0.0923	0.1961	0.133	0.0674	0.3099
	Survival	0.855	0.794	0.882	0.947	0.857
3	Hazard	0.322	0.4615	0.6154	0.3855	0.5976
	Survival	0.776	0.638	0.765	0.883	0.591
4	Hazard	0.487	0.3714	0.1667	0.6222	0.6522
	Survival	0.526	0.344	0.294	0.543	0.238
5	Hazard	0.666	0.5625	0.999	0.334	0.51
	Survival	0.269	0.216	0.245	0.205	0.083

Source: own calculations

Figure 1: Survival



4 References

ANDERSON, P.K., R.D. GILL : Cox's regression model for counting processes: A large sample study. Annals of Statistics, Vol.10, No.4, 1982, pp.1100-1120.

ARJAS, E., A. HAARA : A logistic regression model for hazard: asymptotic results. Scand. J. Stat. 14, 1987, pp.1-18.

ARJAS, E., A. HAARA : A note on the asymptotic normality in the Cox regression model. Annals of Statistics, Vol.16, No.3, 1988, pp.1133-1141.

BRECHT, B. : Aufbau, Struktur und Anwendungen des Sozio-ökonomischen Panels in INGRES. Diskussionsbeitrag Nr.II-120, SFB 178, Universität Konstanz, 1990.

BRECHT, L. : Ein zeitdiskretes Modell zur Verweildaueranalyse: Regularitätsbedingungen und asymptotische Ergebnisse. Diskussionsbeitrag Nr. 127/s, Fakultät für Wirtschaftswissenschaften und Statistik, Universität Konstanz, 1991(a).

BRECHT, L. : Ansätze zur semiparametrischen Regressionsanalyse multivariater korrelierter Verweildauermodelle. Diskussionsbeitrag Nr. 129/s, Fakultät für Wirtschaftswissenschaften und Statistik, Universität Konstanz, 1991(b).

BYE, B.V., G.F. RILEY : Model estimation when observations are not independent. Sociological Methods and Research 17, 1989, pp 353-357.

GILL, R.D. Censoring and stochastic integrals. Mathematical Centre Tracts, Amsterdam, 1980.

GLAMOUR , User's Guide, Version 2.0. Institut für Statistik und Wirtschaftsgeschichte, Universität Regensburg, 1990.

HAMERLE, A., G. TUTZ : Diskrete Modelle zur Analyse von Verweildauer und Lebenszeiten. Campus Verlag, 1989.

PRODISA (Program for Discrete Survival Analysis), Version 1.5. Faculty of Economics and Statistics, PROSA Project Group, University of Konstanz, 1992.

REBOLLEDO, R. Central limit theorems for local martingales. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 51, 1978, 269-286.

5 Appendix

To prove the results of chapter 2 we make use of martingale theory. Some results from martingales are cited here:

Let $M(t)$ be a martingale then it holds

Theorem 5.1 *(Lenglart's inequality)*

$$P(\sup_{t\in[0,T]} \mid M(t) \mid > \eta) \leq \frac{\delta}{\eta^2} + P(\langle M\rangle(T) > \delta), \text{for all} \quad \delta, \eta > 0$$

Proof: see Anderson / Gill (1982).

Definition 5.1 (ϵ-decomposition)
Let $M_i(t), i = 1, \cdots, r$ be local square integrable martingale.
If there exists for all $\epsilon > 0$ local square integrable martingales
$\bar{M}_1^\epsilon, \cdots, \bar{M}_n^\epsilon, \hat{M}_1^\epsilon, \cdots, \hat{M}_n^\epsilon$, so that for all i with $i = 1, \ldots, n$,
(i) $M_i(t) = \bar{M}_i^\epsilon(t) + \hat{M}_i^\epsilon(t)$,
(ii) $\sup_{t\in[0,\infty)} \mid \hat{M}_i^\epsilon(t) - \hat{M}_i^\epsilon(t^-) \mid \leq \epsilon$ almost sure,
(iii) $\bar{M}_i^\epsilon$ has paths of locally bounded variation,
(iv)

$$P(\exists t \in [0,\infty),\ \hat{M}_i^\epsilon(t) - \hat{M}_i^\epsilon(t^-) \neq 0 \quad \bar{M}_i^\epsilon(t) - \bar{M}_i^\epsilon(t^-) \neq 0) = 0,$$

holds, then $\{\bar{M}_1^\epsilon(t), \cdots, \bar{M}_n^\epsilon(t)\}$ is called jump part of an ϵ-decomposition of $\{M_1(t), \cdots, M_n(t)\}$.

Theorem 5.2 *Let $M_i^{(n)}$ be local square integrable martingales, M_i a Gaussian martingale and for all $\epsilon > 0$ and for all n there exists an ϵ-decomposition of $M_i^{(n)}$ with*

(i)

$$\langle \bar{M}_i^{(n)\epsilon}, \bar{M}_i^{(n)\epsilon}, \rangle(t) \xrightarrow{p} 0, \quad n \to \infty \quad \text{for all } i, t$$

(ii)

$$\langle M_i^{(n)}, M_j^{(n)}\rangle(t) \xrightarrow{p} \begin{cases} Var(M_i) & i = j, \\ 0 & \text{otherwise,} \end{cases} \quad \begin{matrix} n \to \infty \\ \text{for all} \quad i, j, t \end{matrix}$$

it then follows

$$M^{(n)}(t) \xrightarrow{d} M(t)$$

Proof: Rebolledo, (1978)