The Dynamics of Transitions between Specific Occupations

Thorsten Heimann
Institute for Econometrics
University of Muenster
Am Stadtgraben 9
48143 Muenster, Germany
E-mail: 05thhe@wiwi.uni-muenster.de

February 2006

Abstract

Previous literature mainly focuses on either transitions from any occupation to any other occupation, or on changes in job status. This paper suggests a model for transitions between specific occupations. We employ a fully parametrised hazard model with competing risks to model the consecutive jumps of an individual between different occupations. The duration spent in a particular occupation follows a log-logistic distribution depending on the age of the individual and the relative earnings level of the destination occupation.

The model is estimated using a maximum likelihood technique and panel data. Drawing a subsample, we apply our model to micro-data from the German social security agency from 1975 to 1995 (IABS). We find that hazard functions are monotonically decreasing and that the influence of age on the transition hazards tends to be negative for many transitions. Relative earnings levels have mixed influence on the hazards.

Keywords: Occupation, Hazard, Log-logistic, Panel data, Estimation

JEL classification: J62, J24, C41

1 Motivation

An individual engaging in a specific occupation bears, inter alia, the risk that future earnings in this occupation may decline over time. The relative

earnings levels of occupations may change for reasons beyond the control of individuals, and a person that has started out in an occupation with prospective earnings opportunities might find himself in a declining occupation years later. As a result, that person either faces lower earnings (or even unemployment), or changes to another occupation. The latter option usually renders useless that part of a person's knowledge and experience that is specific only to a particular occupation.[1] Although this loss of human capital may be offset by a change to an occupation whose overall earnings level has performed relatively better it may as well lead to a substantial decline in relative earnings.

Thus, the earnings of a whole occupational group are at risk. Shiller (2003) sets out this idea in a much broader framework, noting that human capital forms a large portion of individual wealth and thus should be hedged as far as possible. Presently, there are only very limited possibilities to diversify such risks, mainly through insurance against short-term unemployment or against occupational invalidity.

Hedging this earnings risk could be implemented, for example, by offering insurance contracts on an appropriate earnings index,[2] as proposed by Shiller and Schneider (1998), or by issuing financial derivatives with earnings indices as an underlying. Whatever the actual design of the hedging instrument is, it is necessary to quantify the earnings risk. To do this, one needs to model the earnings of a person that starts out in a particular occupation today and then stochastically jumps between occupations in the future. This model of earnings dynamics can then be used to simulate future earnings paths. Since the earnings risk must depend, inter alia, on the risk of changing occupation in the future occupation must enter the model stochastically. The interest of this paper lies in the construction of such a model for occupational choice.

The literature mainly focuses on two questions: One branch of literature asks whether or not a person is employed (i.e., changes in job status), usually with the purpose of modelling labour supply. Other models aim at explaining the risk of changing occupation at all, where an occupational change is defined as a transition from any occupation to any other occupation. However, neither of these models explains changes between specific occupations, which would be needed for quantifying occupational earnings risk. A possible reason for the lack of such models might be that appropriate datasets and feasible estimation techniques have not been available up to a

[1] In contrast, a change of job (i.e., a change of the employer) implies a loss of firm-specific but not necessarily occupation-specific knowledge and experience.

[2] Using an aggregated index is necessary to prevent moral hazard.

more recent time.

The paper is organised as follows. Section 2 sets forth a fully parametrised hazard model with competing risks for the consecutive jumps between specific occupations, and describes an estimation technique. We then specialise to the case of a log-logistic hazard. In section 3 we estimate our model using a unique administrative dataset from Germany. Section 4 concludes.

2 The Model and Estimation Technique

In this section, we set up a hazard model with competing risks to construct a likelihood for a sample of occupational paths. Let the superscript l denote the individual. J_k^l denotes the k-th failure and T_k^l the corresponding failure time (measured from the same initial time origin). The vector $Z_{ij}^l\left(t_k^l\right)$ may include both the covariates $x_{ij}^l\left(t_k^l\right)$ and the history of the process up to the beginning of spell k, $\{(t_r, j_r); r = 1, \ldots, k-1, k > 1\}$. The covariates may include, for example, the individual's age, gender, job position, education, and firm size, as well as occupational earnings. Given state i in interval $k-1$, the hazard that the k-th failure at time t_k^l is of cause $j \neq i$ is defined as

$$\begin{aligned} & \lambda_{ij}\left(t_k^l; Z_{ij}^l\left(t_k^l\right)\right) \\ = & \lim_{\Delta t \to 0} \frac{P\left(t_k^l \leq T_k^l \leq t_k^l + \Delta t, J_k^l = j | T_k^l \geq t; j_{k-1}^l = i; Z_{ij}^l\left(t_k^l\right)\right)}{\Delta t}, \end{aligned} \quad (1)$$

which depends on a set of parameters θ_{ij}. For notational convenience, we drop the dependence on θ_{ij} for the time being. The hazard rate of leaving state i is the sum over the hazards for all m types of possible destination states,

$$\lambda_i\left(t_k^l; \left\{Z_{ij}^l\left(t_k^l\right)\right\}_{j=1, i\neq j}\right) = \sum_{j=1, i\neq j}^{m} \lambda_{ij}\left(t_k^l; Z_{ij}^l\left(t_k^l\right)\right). \quad (2)$$

Following, for example, Kalbfleisch and Prentice (2002) and Lancaster (1990), the likelihood for the observed spells of individual l is given by

$$\begin{aligned} L^l = & \prod_{k=1}^{r_l} \left[\exp\left(-\int_{t_{k-1}^l}^{t_k^l} \lambda_{i_k^l}\left(t; \left\{Z_{i_k^l j}^l\left(t_k^l\right)\right\}_{j=1, i\neq j}\right) dt\right)\right. \\ & \left. \times \left[\lambda_{i_k^l j_k^l}\left(t_k^l; Z_{i_k^l j_k^l}^l\left(t_k^l\right)\right)\right]^{c_k^l}\right], \end{aligned} \quad (3)$$

where c_k^l is 0 if j_k^l is censored and 1 otherwise. We now aim at rewriting this likelihood in order to enable estimation of θ, and define indicator variables

$$\begin{aligned}\delta_{ik}^l &= \begin{cases} 1 & \text{if state } i \text{ is entered at the end of interval } k, \\ 0 & \text{otherwise,} \end{cases} \\ \delta_{i\neq j}^l &= \begin{cases} 1 & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases}\end{aligned}$$

Using these indicators and inserting (2) into (3) yields

$$\begin{aligned}L^l &= \prod_{k=1}^{r^l}\prod_{i=1}^{m}\prod_{j=1}^{m}\left[\exp\left(-\int_{t_{k-1}^l}^{t_k^l}\delta_{i,k-1}^l\delta_{i\neq j}^l\lambda_{ij}\left(t;Z_{ij}^l\left(t\right)\right)dt\right)\right. \\ &\quad \left.\times\left(\lambda_{ij}\left(t_k^l;Z_{ij}^l\left(t_k^l\right)\right)\right)^{\delta_{i,k-1}^l\delta_{jk}^l c_k^l}\right].\end{aligned}$$

As the occupational choice can be assumed independent across the n individuals in the sample, the joint likelihood is

$$L = \prod_{l=1}^{n} L_l,$$

and after some rearranging we obtain the joint log-likelihood as

$$\begin{aligned}LL &= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{l=1}^{n}\sum_{k=1}^{r^l}\left[\delta_{i,k-1}^l\delta_{jk}^l c_k^l\ln\lambda_{ij}\left(t_k^l;Z_{ij}^l\left(t_k^l\right)\right)\right. \\ &\quad \left.-\delta_{i,k-1}^l\delta_{i\neq j}^l\int_{t_{k-1}^l}^{t_k^l}\lambda_{ij}\left(t;Z_{ij}^l\left(t\right)\right)dt\right]. \qquad (4)\end{aligned}$$

This log-likelihood breaks up into separate components for each λ_{ij}. We assume that θ can be partitioned into subsets θ_{ij}, for all feasible i and j, such that for each transition the intensities depend upon disjoint sets of parameters. Then the log-likelihood contributions can be maximised separately for each transition, which greatly facilitates estimation. That is, for given i and j, we obtain the estimates $\hat{\theta}_{ij}$ by numerically maximising

$$\begin{aligned}LLC_{ij} &= \sum_{l=1}^{n}\sum_{k=1}^{r^l}\left[\delta_{i,k-1}^l\delta_{jk}^l c_k^l\ln\lambda_{ij}\left(t_k^l;Z_{ij}^l\left(t_k^l\right)\right)\right. \\ &\quad \left.-\delta_{i,k-1}^l\delta_{i\neq j}^l\int_{t_{k-1}^l}^{t_k^l}\lambda_{ij}\left(t;Z_{ij}^l\left(t\right)\right)dt\right] \qquad (5)\end{aligned}$$

with respect to the elements of θ_{ij}.

We now specify a functional form for the hazard (1). Denoting by $D_k^l = T_k^l - t_{k-1}^l$ the duration of spell k, we set (see, e.g., Blossfeld, Hamerle, and Mayer (1989))

$$\ln D_k^l = \mu_k^l + \sigma\omega_k^l$$

and assume for ω_k^l a zero-mean logistic distribution with density

$$f\left(\omega_k^l\right) = \frac{e^{\omega_k^l}}{\left(1+e^{\omega_k^l}\right)^2}.$$

Then the hazard for failure after elapsed duration d_k^l is of the log-logistic form

$$\lambda\left(d_k^l\right) = \frac{\alpha\left(\gamma_k^l\right)^\alpha\left(d_k^l\right)^{\alpha-1}}{1+\left(\gamma_k^l d_k^l\right)^\alpha},$$

where $\gamma_k^l = e^{-\mu_k^l}$ and $\alpha = \sigma^{-1}$. This hazard is flexible in that it is monotonically decreasing from infinity for $\alpha < 1$, monotonically increasing from γ_k^l for $\alpha = 1$, and takes an inverted U-shaped form for $\alpha > 1$ (increasing from zero to a single maximum and decreasing toward zero thereafter). Further, if the covariates are chosen to be time-invariant there exists a simple closed form for the survivor function, allowing to explicitly solve the integral in (5).

To obtain a regression model, we let the mean duration depend on $Z^l(t)$ (including a constant term) and parametrise $\gamma_k^l = \exp\left(Z^l\left(t_k^l\right)'\beta\right)$, with coefficient vector β. Additionally substituting $d_k^l = t_k^l - t_{k-1}^l$, we take (see also Lancaster (1990))

$$\lambda_{ij}\left(t_k^l; Z_{ij}^l\left(t_k^l\right)\right) = \frac{\alpha_{ij}\exp\left(\alpha_{ij}Z_{ij}^l\left(t_k^l\right)'\beta_{ij}\right)\left(t_k^l - t_{k-1}^l\right)^{\alpha_{ij}-1}}{1+\left(\exp\left(Z_{ij}^l\left(t_k^l\right)'\beta_{ij}\right)\left(t_k^l - t_{k-1}^l\right)\right)^{\alpha_{ij}}} \tag{6}$$

as our hazard function. The set of parameters is $\theta_{ij} = \left\{\alpha_{ij}, \beta_{ij}\right\}$, where α_{ij} is a scalar volatility parameter and β_{ij} a coefficient vector. With the hazard

function specified as in (6), the log-likelihood contribution (5) becomes

$$\begin{aligned} & LLC_{ij} \\ = & \sum_{l=1}^{n}\sum_{k=1}^{r^l}\left[\delta_{i,k-1}^{l}\delta_{jk}^{l}c_{k}^{l}\left[\alpha_{ij}Z_{ij}^{l}\left(t_{k}^{l}\right)'\beta_{ij}+\ln\left(\alpha_{ij}\left(t_{k}^{l}-t_{k-1}^{l}\right)^{\alpha_{ij}-1}\right)\right.\right. \\ & \left.-\ln\left(1+\exp\left(\alpha_{ij}Z_{ij}^{l}\left(t_{k}^{l}\right)'\beta_{ij}\right)\left(t_{k}^{l}-t_{k-1}^{l}\right)^{\alpha_{ij}}\right)\right] \\ & \left.-\delta_{i,k-1}^{l}\delta_{i\neq j}^{l}\int_{t_{k-1}^{l}}^{t_{k}^{l}}\frac{\alpha_{ij}\exp\left(\alpha_{ij}Z_{ij}^{l}\left(t\right)'\beta_{ij}\right)\left(t-t_{k-1}^{l}\right)^{\alpha_{ij}-1}}{1+\exp\left(\alpha_{ij}Z_{ij}^{l}\left(t\right)'\beta_{ij}\right)\left(t-t_{k-1}^{l}\right)^{\alpha_{ij}}}dt\right]. \end{aligned} \tag{7}$$

An estimator for the covariance matrix of the parameter estimator $\hat{\theta}$ can be obtained in the standard manner. As the observed occupational paths can be considered to be iid across individuals their ML estimators are consistent and asymptotically normal with mean θ and covariance matrix

$$\mathbf{V}\left(\theta\right)=\mathcal{I}\left(\theta\right)^{-1}, \tag{8}$$

where $\mathcal{I}\left(\theta\right)=-\operatorname{E}\left(\mathbf{H}\left(\theta\right)\right)$ is the Fisher information and $\mathbf{H}\left(\theta\right)$ is the Hessian of the log-likelihood function. Since (8) contains the unknown parameter set θ we take as an estimator for the asymptotic covariance matrix of $\hat{\theta}$ the inverse of the negative Hessian evaluated at $\hat{\theta}$,

$$\hat{\mathbf{V}}\left(\hat{\theta}\right)=-\mathbf{H}\left(\hat{\theta}\right)^{-1}. \tag{9}$$

Recall that the Hessian $\mathbf{H}\left(\theta\right)$ has (p,q) matrix entries

$$\mathbf{H}_{pq}\left(\theta\right)=\frac{\partial^{2}LL\left(\theta\right)}{\partial\theta_{p}\partial\theta_{q}}$$

for transitions (i_p,j_p) and (i_q,j_q). Since the parameter set θ is assumed to consist of disjoint subsets θ_{ij} the log-likelihood sums into separate terms for each θ_{ij} and it follows that the cross derivatives (i.e., $p\neq q$) are all zero. Thus $\mathbf{H}\left(\hat{\theta}\right)$ is block diagonal with the Hessians of the log-likelihood contributions of each transition (i,j) with respect to the elements of θ_{ij} as block diagonal entries.

Since, generally, the inverse of a block diagonal matrix is obtained by simply inverting the block diagonal (matrix) elements, $\hat{\mathbf{V}}\left(\hat{\theta}\right)$ is also block diagonal and the covariance matrices of the θ_{ij} can be computed separately for each transition (i,j).

3 Data and Results

Estimation of our model is based on the Institute for Employment Research (Institut für Arbeitsmarkt und Berufsforschung, IAB) employment subsample, which covers a one percent random sample of all employees registered with the German social security system within the period from 1975 to 1995. It includes spell data on the employment history as recorded by the social insurance system, and information on periods of drawing unemployment benefits. The variables include information on education, part/full time employment, occupation and the average daily remuneration during the spell. In addition, some socio-economic variables such as age, gender, marital status, nationality, number of children, etc. are available.

The data have been recorded by the administrative data collection procedure of the social insurance system, introduced in West Germany in 1973. It includes a common notification procedure for health insurance, unemployment insurance, and the statutory pension scheme. All employers in Germany are legally obliged to supply the social security agencies with comprehensive information about their employees. Employers have to notify the agencies of any relevant changes in the employment status. If there are no changes an annual control notification is required. These data are collected and stored by the Federal Employment Service (Bundesanstalt für Arbeit). Since each change in employment status triggers a notification the information is constantly updated, and each time a new spell is created. The beginning of a new spell therefore does not necessarily imply another occupational status. Further, the spell lengths vary according to the frequency of the notifications submitted by the employer. For further information on the dataset see Bender, Haas, and Klose (2000).

Since the purpose of the data collection is to set up a social insurance account for each employee, and since substantial legal sanctions are imposed for incorrect or missing notifications, the data are much more reliable than survey data collected on a voluntary basis. Furthermore, the dataset does not suffer from panel mortality or attrition.

However, the German social insurance system does not include civil servants, self-employed persons, nor employees with earnings below a certain threshold and therefore not subject to social insurance contributions. In 1995, the employees registered with the social insurance system in West Germany accounted for roughly 80 percent of the total workforce, varying across occupations and industries (see Bender, Haas, and Klose (2000)). Further, the social insurance agency records a person's earnings only up to the contribution assessment ceiling of the social security system. Earnings

$i \backslash j$	1	2	3	4	5	6	7	8	9	10	11
1		0.63	0.86	0.88	0.99	0.82	0.60	0.70	0.61	0.76	1.09
2	0.76		0.93	0.97	0.86	0.50	0.72	0.58	0.63	0.68	1.31
3	0.82	2.14		0.85	0.56	0.69	0.36	1.00	0.99	0.68	1.68
4	0.81	0.81	0.73		0.90	0.56	0.83	0.77	0.54	0.78	0.95
5	1.09	0.98	1.15	1.01		0.76	0.70	0.55	1.70	0.98	1.22
6	0.81	0.66	0.69	0.65	0.79		0.68	0.75	0.66	0.80	1.00
7	0.90	0.75	0.89	0.89	1.24	0.65		0.69	0.63	0.62	1.04
8	0.72	0.81	0.81	0.92	0.93	0.57	0.68		0.66	0.70	1.06
9	0.84	0.46	0.58	0.79	0.62	0.75	0.62	0.74		0.75	1.57
10	0.75	0.70	0.77	0.83	0.72	0.71	0.59	0.67	0.63		1.05

Table 1: Estimation results for the α parameters

$i \backslash j$	1	2	3	4	5	6	7	8	9	10	11
1		−41	28	56	25	56	−48	−51	−12	−64	−4
2	20		26	56	25	11	−82	−78	−2	−65	−8
3	−31	−34		−38	83	−54	−909	−16	−2	−28	−7
4	−47	−56	41		36	2	−43	−40	−18	−36	−3
5	−28	−21	−36	−44		−42	−31	−21	−10	−18	−5
6	−35	−14	11	−4	11		−33	−32	−11	−29	−4
7	37	76	17	44	14	30		−57	−16	−18	−6
8	39	40	20	31	24	30	38		−13	50	−5
9	7	3	11	9	2	7	5	14		10	−4
10	55	53	13	36	20	40	23	−60	−15		−5

Table 2: Estimation results for the β_0 parameters (intercept)

exceeding this threshold are censored (from above). The threshold increases over time roughly in line with the increase in the general wage level. For the 1980s, the fraction of censored observations lies between 8 and 11 percent, but it is substantially higher for subgroups such as highly qualified employees (see Steiner and Wagner (1997)).

The complete IAB dataset contains 7 847 553 observations (i.e., spells) on 559 540 individuals. In order to reduce the computational requirements and to facilitate the statistical analysis, we exclude individuals from the original dataset in the following way.

First, we remove persons with missing occupational information or with obviously implausible observations, especially spells with zero length and daily earnings of less than one or more than 300 DM (far above the assessment ceiling). Individuals with only one observation are eliminated, as well as apprentices, trainees, home workers, part-time workers, and people with

$i \diagdown j$	1	2	3	4	5	6	7	8	9	10	11
1		−2.1	−1.4	−0.2	−1.2	−1.8	−1.6	−1.6	−1.3	−1.3	3.3
2	−0.4		−1.5	−0.6	−1.2	−0.1	0.1	−0.0	−1.4	0.1	5.7
3	−0.2	−1.3		−0.6	1.1	−6.4	−15.0	−1.6	−7.6	0.0	5.2
4	−0.8	−1.3	−2.6		−0.7	−1.4	−1.3	−1.5	−0.8	−1.3	2.8
5	−0.6	−0.9	−1.8	0.2		−1.4	0.8	−2.9	0.1	−1.8	4.2
6	−2.1	−1.8	−2.5	−0.5	−2.1		−0.5	−2.2	−1.4	−0.8	3.2
7	−1.5	−0.5	−6.2	−0.3	−0.5	−0.8		−0.8	−1.0	−0.7	4.3
8	−0.6	−1.1	−4.3	−0.6	−1.9	−0.1	−1.3		−1.8	−1.3	4.0
9	−0.9	−1.6	−3.3	−0.6	0.4	0.2	−2.3	−2.2		−0.6	4.0
10	−1.2	−1.7	−1.4	−0.9	−2.0	−0.4	−1.1	−1.5	−1.4		3.8

Table 3: Estimation results for the β_1 parameters (age)

$i \diagdown j$	1	2	3	4	5	6	7	8	9	10
1		39	−29	−55	−21	−64	45	52	14	74
2	−21		−26	−53	−22	−19	80	78	−3	72
3	30	35		37	−79	62	861	17	10	31
4	49	57	−42		−33	−5	44	43	20	44
5	33	24	41	49		51	31	22	14	24
6	31	11	−12	2	−10		28	31	12	31
7	−37	−78	−15	−42	−12	−35		60	19	21
8	−39	−39	−20	−28	−20	−35	−36		17	−51
9	−4	−6	−9	−6	−5	−6	−4	−10		−6
10	−49	−48	−14	−30	−17	−40	−22	56	18	

Table 4: Estimation results for the β_2 parameters (relative earnings)

unspecified or unknown status. Further, we restrict our analysis to male employees that work in West Germany (as the observation period for East German employees is too short). Finally, we drop individuals whose daily earnings exceed the contribution assessment ceiling.

The model of the last section allows transitions only between different occupations (i.e., from i to $j \neq i$) so that the end of each spell always comes along with a change in occupational status. Since in the original IAB dataset this is frequently not the case the model would be heavily misspecified for these data. We therefore merge consecutive spells of the same individual if the origin and destination states are identical. If these spells are not adjacent, periods between the spells are simply included in the merged spell. If someone stops working and later restarts work in the same occupation this procedure does not treat it as an occupational change. We then delete spells starting at the beginning of the observation period[3] since most of them are possibly left-censored.

Drawing this subsample from the original IAB data and merging some spells reduces our dataset to 23 437 observations on 11 936 individuals, which is the final dataset used for estimation.[4] Each observation contains an identification number for the individual, the beginning and end of the spell, the person's date of birth, the occupation of the current spell, and total earnings during the spell. For all individuals and spells, the mean length of all spells is 3.86 years and the median 1.93 years, and the mean daily earnings are 89.65 DM.

The data provider has defined the occupations in the dataset in accordance with a detailed 4-digit code list from the Federal Employment Service (Berufsklassifikation der Bundesanstalt für Arbeit), which refers to industries and job contents (see Bender, Haas, and Klose (2000)). In order to anonymise the dataset, the IAB has cut these codes to three digits and merged some groups. Still, there are 243 occupational codes in the original dataset and the number of parameters is too large for estimation (since the number of transitions in the model is quadratic in the number of occupations).

In order to reduce the number of occupations we cluster the occupational codes based on the transition matrix. This matrix contains for each (i, j)

[3] To be precise, we delete spells beginning one month later (before 1 February 1975) because the data provider has shifted dates randomly along the time axis in order to anonymise the data.

[4] As the actual estimation progress is quite time-consuming the current results (except the clustering) are based on a random 15-percent subsample of the dataset. An estimation using the full dataset is in progress.

$i \diagdown j$	1	2	3	4	5	6	7	8	9	10	11
1		1.4	0.0	0.0	1.0	0.0	1.8	11.2	5.0	20.9	1.0
2	0.1		0.0	0.0	1.0	0.0	17.4	14.8	0.0	18.6	0.2
3	4.9	18.9		7.8	0.6	16.9	9.8	2.2	0.9	15.5	0.1
4	18.7	18.2	0.0		0.5	0.2	16.6	18.6	8.3	21.2	1.2
5	29.6	18.7	30.6	27.7		20.9	11.4	1.7	45.6	26.2	0.6
6	0.1	0.1	0.0	0.6	0.3		0.4	0.8	3.2	16.3	1.6
7	0.0	0.0	0.0	0.1	0.4	0.0		18.0	8.8	10.8	0.6
8	0.1	0.0	0.0	0.5	0.5	0.0	0.2		6.6	0.0	0.7
9	2.5	0.2	0.3	5.2	0.4	2.0	0.5	2.8		6.7	1.1
10	1.1	0.3	0.1	3.6	1.0	0.1	0.6	2.0	8.4		1.0

Table 5: Hazards rates for a transition from i to j after one year for a 35-year-old individual

the fraction of people in i changing to j. We then compute for each pair the mean of the transition fractions for both directions, i.e. from i to j and from j to i. As distance measure between i and j we take one minus this mean fraction, so that two occupations are the closer the larger the fraction of people changing between them. The resulting clusters are thus homogeneous in a way that many people change within them but only relatively few between them. We employ a hierarchical clustering algorithm and then cut the dendrogram such that exactly ten occupational groups are produced (with codes "1" to "10"). Table 6 shows the grouping. Details about the grouping procedure can be found in the appendix.

Group	Contents
1	Business, Organisation, Management
2	Electricians
3	Health, Medical professions
4	Mining, Industry (partly), Administrative tasks, Arts, Teaching
5	Engineering, Natural Sciences, Other Academics, Management consultants
6	Construction
7	Locksmiths
8	Metalproducing and metalworking industry
9	Hotel and restaurant industry
10	Agriculture, Industry (partly), Food Industry, Craft

Table 6: Grouping of occupations

As time unit we chose days elapsed since 1 January 1900. Since only the year of birth is given in the data we set each person's day of birth to 30 June.

Generally, for the computation of the likelihood it is necessary to know the occupation of each following spell, which is though not available for a person's last observation. We treat this problem in two different ways, according to the two fundamentally different mechanisms behind the censoring. First, we consider all spells ending at the end of the observation period[5] as randomly right-censored and set their censoring indicators c_k^l equal to 0. Second, we assume that, after their last spells, the remaining persons definitely exit the study group (i.e., the work force subject to social security contributions, and people drawing unemployment benefits). This second group of people leaving the dataset presumably consists of people who retire, become self-employed, become civil servants, or completely drop out of the labour market. As a substantial portion of the dataset (11 936 spells) is affected and simply dropping these spells would clearly bias the results we incorporate this group into our model by introducing a new, separate occupational status with code "11".

For each employee, the dataset contains socio-economic variables on the person as well as information on their establishment. However, although the parameters are estimated separately for each transition numerical maximisation of the likelihood becomes tedious if many covariates are included. In order to facilitate estimation and also to keep the model simple, we include only two covariates: the individual's age and the earnings in the subsequent occupation (j) relative to the earnings in the current occupation (i). While age is allowed to vary relative earnings are set constant throughout each spell. We compute the relative earnings as the average of the daily earnings of all individuals in occupation j at the last day of the spell divided by the corresponding earnings of the individuals in occupation i. This is done using the dataset before the spells are merged. For changes to state 11 we exclude relative earnings from the covariates since we do not observe the earnings level in state 11 and, moreover, most of the people in this group are probably retirees whose retirement decision is predominantly governed by their age.

The parameter estimates are obtained by numerically maximising the log-likelihood contributions (7) separately for each transition type.[6] The

[5] Again, we account for the anonymisation by inserting an extra month and include all spells ending after 1 December 1995.

[6] Maximisation was performed with the `ConstrOptim` command of the statistical programming language `R`, version 2.2.1, which uses the `optim` command implementing the method of Nelder and Mead (1965). Appropriate linear restrictions have been placed on the parameter range to avoid numerical problems which occur when the optimisation algorithm tries too large parameter values. None of the optima is near the boundaries.

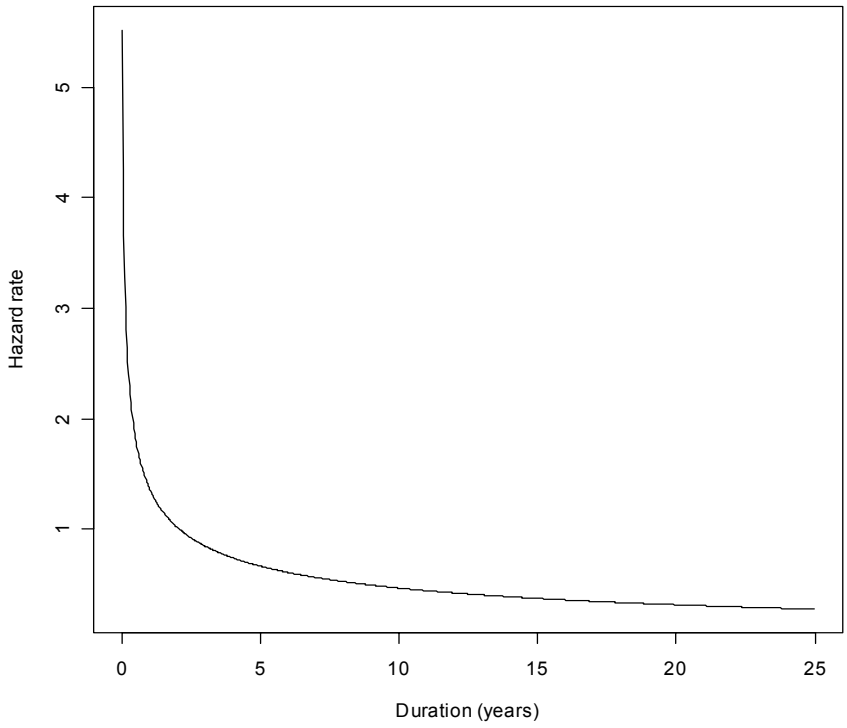

Figure 1: Hazard function for transition from occupational groups 1 to 2 for a 35-year-old individual

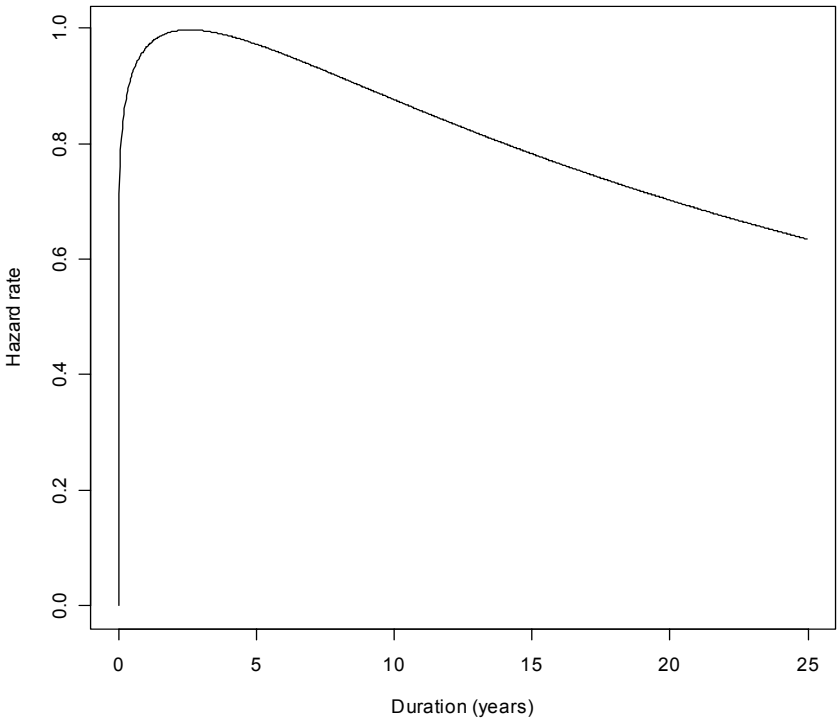

Figure 2: Hazard function for transition from occupational groups 1 to 11 for a 35-year-old individual

results for the αs are given in table 1, for the intercepts β_0 in table 2, for the age coefficients β_1 in table 3, and for the relative earnings coefficients β_2 in table 4. We have not estimated the diagonal elements (since the model rules out these transitions), nor the parameters for leaving state 11 (since it is absorbing), nor β_2 for changing to state 11 (since it does not exist in those hazards).

Except for changes to state 11, the estimates for the α parameters are almost all smaller than one, which suggests that the hazard functions are monotonically decreasing in the vast majority of cases. For changes to state 11, α is larger than 1 and the hazard follows an inverted U-shaped form. Since testing for significance does not make sense for the αs we calculate the widths of 99-percent confidence intervals relative to the point estimates, which have a mean of 66.2 percent and a median of 44.7 percent (complete table in the appendix).

With 81 negative and 9 positive signs, the estimates for the β_1 parameters are predominantly negative for changes to groups 1 to 10. Many estimates are highly significant (p-values reported in the appendix). This suggests that people generally become more reluctant to change their occupation when growing older, while in few, presumably physically more demanding occupations they prefer to leave them as their age increases. As could be expected, the estimates for changes to 11 are all positive as these people are very likely to retire.

The β_2 estimates have mixed signs, 46 negative and 44 positive ones. Many of the estimates are highly significant (p-values reported in the appendix). A positive estimate means that these transitions are more probable if the earnings level in the destination occupation is higher than in the original occupation. Correspondingly, transitions with negative estimates mean that people change to occupations with relatively lower earnings levels. These mixed results are rather counterintuitive, especially the negative coefficients for the expectedly well paid groups 3 and 5. However, a reason may be the exclusion of people with earnings above the assessment ceiling, which produces a downward bias in the relative earnings of the well paid groups. Modifications of the model to account for this bias are subject to current research.

Table 5 sets forth a matrix of the hazard rates for a transition from occupations i to j after one year, with the hazards calculated for a 35-year-old individual and the simplifying assumption that relative earnings are 1.1 for all transitions. Figure 1 shows a plot of the hazard function for a transition from groups 1 to 2 for the same covariates. As $\alpha_{1,2} < 1$ the hazard function decreases monotonically from infinity at the origin to zero

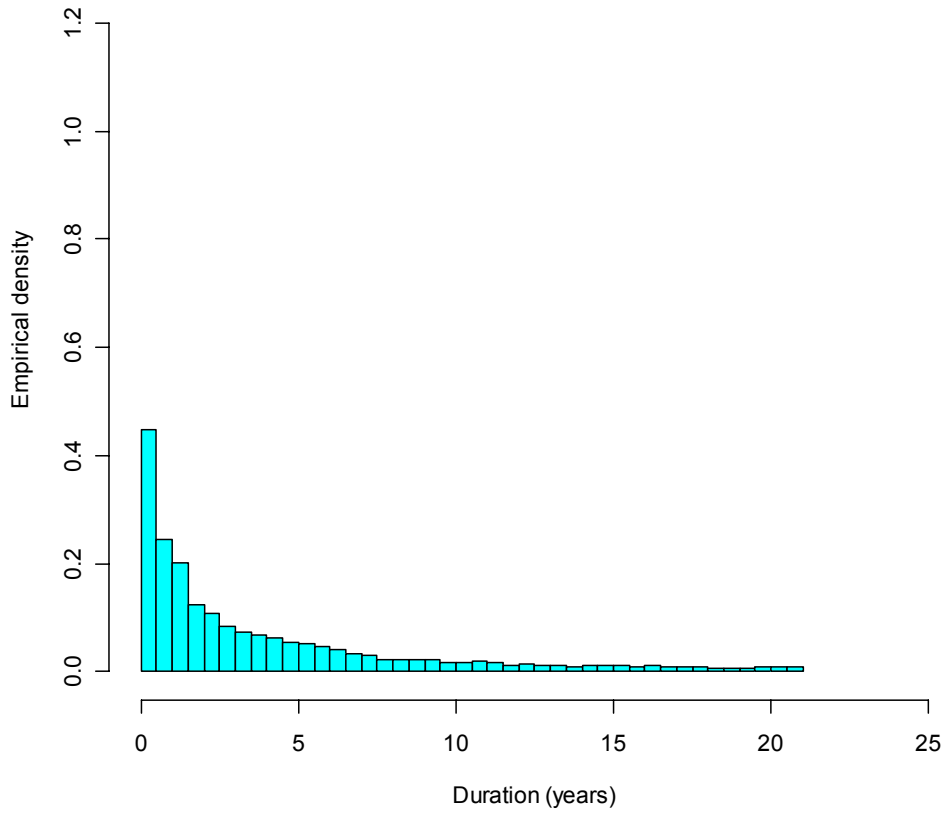

Figure 3: Distribution of the durations in the IAB data used for estimation

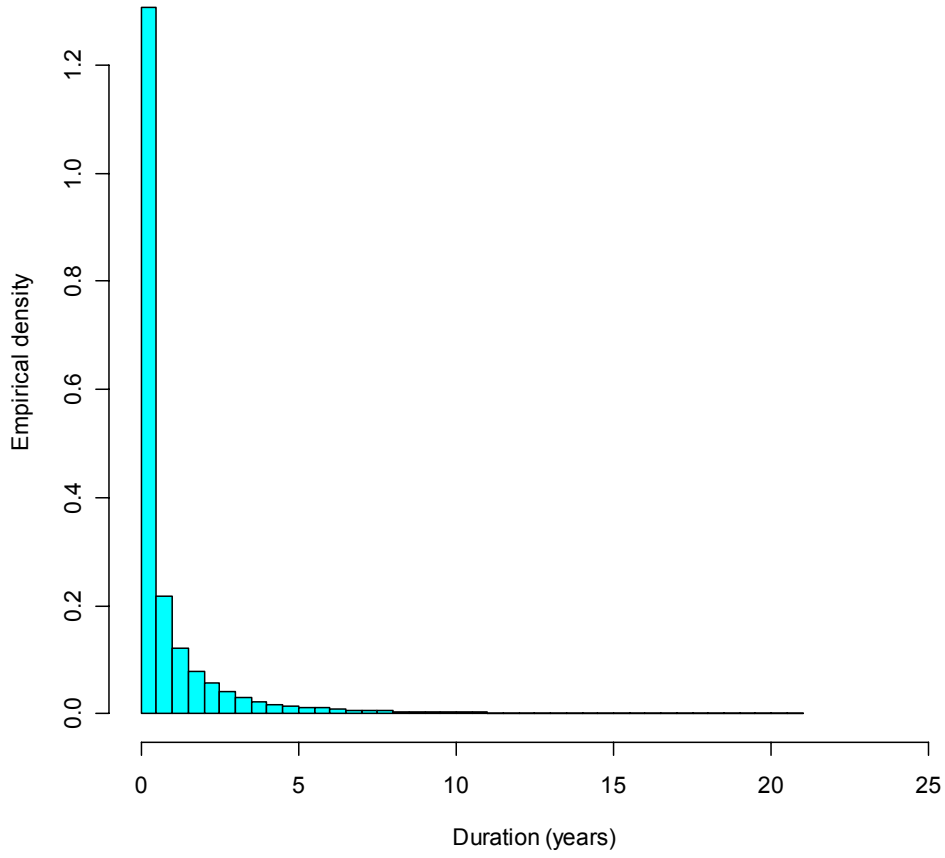

Figure 4: Distribution of the simulated durations

as the duration goes to infinity. In contrast, the hazard in figure 2 for a transition from 1 to 11 takes an inverted U-shape since $\alpha_{1,11} > 1$.

Since a purpose of the model is to simulate future occupational paths we have simulated paths for 10 000 individuals based on the estimation results to see how good model implications fit the data. Figures 3 and 4 show the duration distributions in the dataset used for estimation and in the simulated results respectively. The simulated durations have a conditional[7] mean of 0.97 years and a median of 0.16 years, which is not close to the empirical values. However, since important parts of the estimation procedure are still preliminary (especially the treatment of the assessment ceiling) this should not be considered the final result.

4 Conclusions

We employ a fully parametrised hazard model with competing risks for the successive jumps of an individual between different occupations. The duration spent in a particular occupation follows a log-logistic distribution depending on the age of the individual and the relative earnings level of the destination occupation. The model is estimated using a maximum likelihood technique and panel data. In order to facilitate the estimation, we separate the likelihood and maximise the likelihood contributions separately for each transition.

Drawing a subsample, we apply our model to a unique administrative dataset from Germany, which is rich enough to identify the large number of parameters. Our findings suggest that the hazard function is monotonically decreasing in the vast majority of cases. The influence of age on the transition hazards tends to be negative for most transitions, only for some transitions the hazard rises with age. This suggests that people generally become more reluctant to change their occupation when growing older. Relative earnings levels have mixed influence on the transition hazards.

[7] Spells in the IABS cannot exceed 21 years (the overall observation period), so the empirical distribution is truncated. We therefore also truncate the simulated distribution to enable comparison.

Appendix

Occupational Groupings

In order to prevent small clusters with only very few occupations, prior to the cluster analysis some relatively close occupations have been merged as presented in table 7. Ward's (1963) hierarchical clustering algorithm was then applied to these occupations.[8] This linkage method minimises the variance within clusters and tends to build clusters of equal size (at least rather than other standard methods). The dendrogram was cut such that exactly ten clusters are produced. The original codes assigned to these clusters are set forth in table 8, and figure 5 depicts the numbers of observations in each group in the dataset used for estimation.

Old occupation	**Added to occupation**
072	071
391	392
855	844
722, 723	721
862, 864	861
682, 684	681
352	351
911, 912	411
172, 173, 174, 176	171
31, 41, 44, 51, 53	11
913	411
782	781
502, 503	501
842	841
854	853
546	545

Table 7: Merging of occupations prior to cluster analysis

[8] The clustering was performed using the `hclust` command of the statistical programming package `R`, version 2.2.1.

Group	Contents (IABS codes)
1	681, 683, 687, 693, 694, 751, 753, 771, 772, 781
2	311, 312, 313, 314, 315, 622, 731
3	611, 685, 841, 844, 851, 853, 856, 861, 876
4	71, 91, 141, 142, 171, 282, 284, 285, 302, 303, 321, 351, 354, 356, 371, 372, 373, 374, 376, 481, 491, 492, 521, 541, 542, 547, 548, 625, 626, 627, 628, 629, 632, 633, 634, 691, 692, 701, 702, 703, 704, 706, 711, 713, 716, 721, 724, 726, 733, 763, 773, 783, 784, 791, 792, 793, 801, 803, 811, 821, 822, 823, 831, 832, 833, 834, 835, 836, 837, 857, 863, 872, 873, 874, 875, 877, 891, 893, 901
5	601, 602, 603, 604, 605, 606, 607, 612, 621, 623, 624, 631, 635, 752, 761, 774, 805, 871, 881, 882, 883
6	441, 442, 451, 461, 462, 464, 465, 470
7	241, 251, 252, 261, 262, 263, 270, 271, 272, 274, 275, 283
8	191, 192, 193, 201, 202, 203, 211, 212, 213, 221, 222, 224, 225, 226, 231, 233, 234, 235, 281, 291, 301, 322, 323, 353, 549, 921, 931, 932
9	411, 794, 922, 933
10	11, 61, 81, 101, 111, 112, 121, 131, 132, 135, 143, 151, 161, 162, 163, 164, 175, 177, 181, 182, 242, 304, 305, 331, 332, 341, 343, 361, 392, 401, 402, 403, 412, 421, 422, 424, 431, 432, 433, 452, 453, 463, 482, 483, 485, 486, 501, 511, 512, 514, 522, 531, 544, 545, 686, 712, 714, 741, 742, 743, 744, 934, 935, 936, 937

Table 8: Grouping of the occupations in the IABS (result of cluster analysis)

p-values and Confidence Intervals for the Parameter Estimates

Missing elements (apart from the diagonal) are a result of numerical problems when calculating the standard deviation.

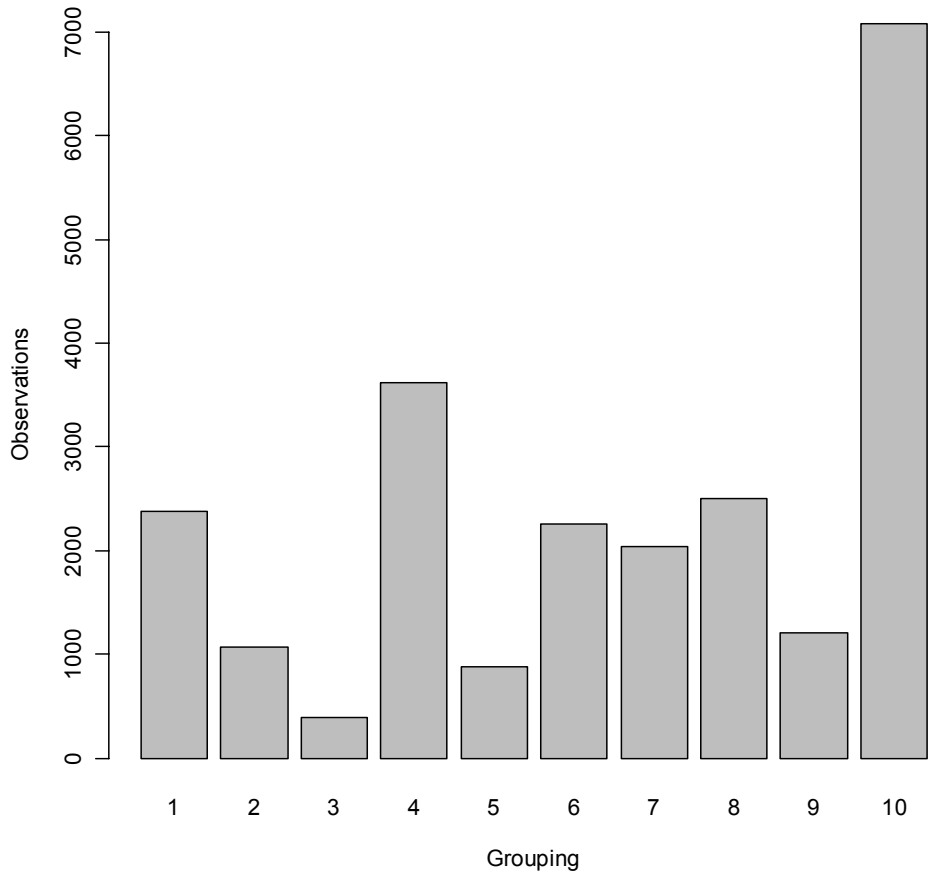

Figure 5: Numbers of observations in the occupational groups

$i \backslash j$	1	2	3	4	5	6	7	8	9	10	11
1		87		10		69	96		67		0
2	55		131	36	65	11	56	40	160	55	
3	123	215		68		161	336	162	211	84	31
4	25	42	7			39	35	27	58	17	36
5	51	102	120			176	133		223		
6	7	84	102	37	17		42	41	3	15	
7		51	151	28		45		26	9	21	
8	59		120		79	45	27		63	16	22
9		0	156	46	162	52	81	4		25	32
10	19			15		16	20	16	28		25

Table 9: Widths of 99-percent confidence intervals relative to the point estimates for the α parameters (in percent)

$i \backslash j$	1	2	3	4	5	6	7	8	9	10	11
1		0.19				0.00	0.16	0.00	0.00		0.00
2	0.12				0.00				0.39	0.03	
3	0.00					0.00		0.04	0.38	0.00	0.00
4			0.00		0.00	0.38	0.00		0.00	0.00	0.00
5	0.00	0.00		0.00		0.03	0.01	0.07	0.04		
6	0.00	0.15		0.34			0.00	0.00		0.00	0.00
7			0.03			0.00				0.12	0.00
8	0.00				0.00	0.00			0.00		0.00
9	0.00		0.06	0.00	0.37	0.00	0.19			0.00	0.00
10						0.00			0.00		0.00

Table 10: p-values for the β_0 parameters

$i \backslash j$	1	2	3	4	5	6	7	8	9	10	11
1		0.02	0.00	0.27		0.01	0.06	0.00	0.06	0.00	0.00
2	0.29		0.23	0.08	0.07		0.39	0.40	0.31	0.38	
3	0.39	0.24		0.22	0.00	0.01	0.33	0.20	0.07	0.40	0.00
4	0.00	0.00	0.06		0.00	0.00	0.00	0.00	0.16	0.00	0.00
5	0.11	0.19	0.12	0.35		0.27	0.30	0.10	0.40		
6	0.00	0.04	0.09	0.12	0.00		0.14	0.00	0.00	0.00	0.00
7		0.23	0.04	0.19		0.10		0.01	0.00	0.01	0.00
8	0.19	0.00	0.02	0.00	0.01	0.38	0.00		0.01	0.00	0.00
9	0.03		0.15	0.13	0.38	0.36	0.04	0.00		0.04	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00

Table 11: p-values for the β_1 parameters

$i \diagdown j$	1	2	3	4	5	6	7	8	9	10
1		0.20	0.00			0.00	0.18	0.00	0.00	
2	0.11		0.00		0.00				0.39	0.02
3	0.00					0.00		0.02	0.11	0.00
4			0.00		0.00	0.34	0.00		0.00	0.00
5	0.00	0.00		0.00		0.03	0.01	0.09	0.01	
6	0.00	0.19		0.37			0.00	0.00	0.00	0.00
7			0.07			0.00				0.10
8	0.00				0.00	0.00			0.00	
9	0.00		0.04	0.00	0.09	0.00	0.13			0.00
10						0.00			0.00	

Table 12: p-values for the β_2 parameters

References

Bender, S., A. Haas, and C. Klose (2000): "The IAB Employment Subsample 1975-1995," *Schmollers Jahrbuch*, 120, 649–662.

Blossfeld, H.-P., A. Hamerle, and K. Mayer (1989): *Event History Analysis: Statistical Theory and Application in the Social Sciences*. Lawrence Erlbaum Associates, Hilldale.

Kalbfleisch, J., and R. Prentice (2002): *The Statistical Analysis of Failure Time Data*. John Wiley and Sons, New Jersey.

Lancaster, T. (1990): *The Econometric Analysis of Transition Data*. Cambridge University Press, Cambridge et al.

Nelder, J. A., and R. Mead (1965): "A Simplex Algorithm for Function Minimization," *Computer Journal*, 7, 308–313.

Shiller, R. (2003): *The New Financial Order: Risk in the 21st Century*. Princeton University Press, Princeton.

Shiller, R., and R. Schneider (1998): "Labor Income Indices Designed for Use in Contracts Promoting Income Risk Management," *Review of Income an Wealth*, 44(2), 163–182.

Steiner, V., and K. Wagner (1997): "Entwicklung der Ungleichheit der Erwerbseinkommen in Westdeutschland," *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung, Institut für Arbeitsmarkt- und Berufsforschung*, 30, 638–641.

Ward, J. (1963): "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 236–244.