Nationaløkonomisk Tidsskrift 144 (2006): 209-229

A Danish Profiling System

Michael Rosholm
Department of Economics, University of Aarhus, *E-mail: mrosholm@econ.au.dk*

Jonas Staghøj
Department of Economics, University of Aarhus, *E-mail: jstaghoej@econ.au.dk*

Michael Svarer
Department of Economics, University of Aarhus, *E-mail: msvarer@econ.au.dk*

Bo Hammer
Mploy A/S, *E-mail: bha@mploy.dk*

SUMMARY: *This paper describes the statistical model used for profiling new unemployed workers in Denmark. When a worker – during his or her first six months in unemployment – enters the employment office for the first time, this model predicts whether or not he or she will be unemployed for more than six months from that date. The case worker's assessment of how to treat the person is partially based upon this prediction.*

»Across OECD countries, millions of unemployed have been out of work for more than a year. And others are at risk of becoming so. One possible way to combat the drift into long-term unemployment is to offer more assistance to job losers before they reach the stage of long-term unemployment. But it would be very costly to offer in-depth help to all of the job losers. This has led some countries to develop methods to both identify jobseekers at risk of becoming long-term unemployed and refer them to suitable labour market programmes, usually known as profiling. But is it possible to accurately identify such jobseekers?«, OECD (1998).

In this paper we present the first statistical component of such a profiling model, which, as of 1 December, 2004, has become an integrated part of the Danish national labour market policy following extensive experimentation with the statistical model and with the way the information is presented to caseworkers. When discussing how to

---

We are grateful to Annette B. Andersen for reading the manuscript and to Klaus Langager from the Danish National Labour Market Authority for access to the data used in the paper. Michael Svarer thanks the Danish National Research Foundation for support through its grant to CAM. Michael Rosholm thanks the Danish Ministry of Social affairs for financial support through the Danish Graduate School for Integration, Production and Welfare.

develop and implement a profiling system, a first requirement is to define which goals the system should be designed to fulfill, because this obviously influences the way the system should be designed. Two natural goals are equity and efficiency and hence it would be interesting to investigate whether these are correlated or in fact conflicting goals however, we will not go into a deep discussion of this issue in the present paper, although we will propose some arguments indicating that both goals may be fulfilled in the present project. After properly defining the goals the next task is to choose a profiling variable based upon which the group of unemployed workers can be divided into different categories. A measure of the probability of becoming long-term unemployed is chosen and the main focus of this paper is on the estimation of this probability and on the predictive power of the probability when using it in the profiling system.[1]

The main purpose of the paper is to verify whether the profiling system is capable of identifying unemployed workers who are at risk of ending up in long-term unemployment (LTU, henceforth). The profiling model consists of a statistical model (presented here) to be used as an initial screening device for identifying potentially long-term unemployed workers, combined with in-depth interviews by caseworkers with those asserted to have a high risk of LTU. The intention is to extend the profiling model by a statistical model and additional interviews designed to identify the 'best' strategy and optimal timing for helping a given unemployed person at risk of LTU in order to reduce the risk of individual LTU.[2]

The statistical component of the profiling system consists of a duration model for the time spent in unemployment. The model is estimated using 120 subgroups, stratified according to age, gender, benefit eligibility, and region of residence. The data used for estimation is the entire inflow into unemployment in Denmark during the period January 1999 – June 2003. Based on the estimated models, it is possible to calculate the probability that a worker attending a meeting with a caseworker at the employment agency will still be unemployed six months from that date, conditional on the elapsed duration of unemployment. A set of threshold values are then calculated in order to maximize the number of correct predictions of the model, and the caseworkers are presented with information about whether the calculated probability is far above, far below, or close to the threshold value.

In several countries attempts have been made to specify worker profiling models. Frölich et al. (2003) state that profiling models are currently used or being tested in Australia, Finland, France, Germany, Ireland, South Korea, New Zealand, Sweden, and the United States. The predictive power of the various models is mixed and to some extent

1. See Black et al. (2000) for a discussion of how to design and evaluate profiling systems.
2. See Frölich et al. (2003) for a discussion of statistically assisted programme selection.

discouraging. Nevertheless, worker profiling is now used in certain states in the U.S. and in South Korea and is, as mentioned above, used on a large scale in Denmark.

Our model is readily comparable to the New Zealand worker profiling model, Watson et al., (1997). The New Zealand model was in active use for some time, but has since been abandoned, allegedly due to a new government that wanted to shift attention from active policies towards incentive-based policies (benefit cuts). Compared to the New Zealand Worker Profiling model, the Danish model provides a substantial improvement in predictive power.

The paper is organized in the following way: Section 2 offers a brief overview of Danish labour market institutions and a recently implemented labour market reform, of which the profiling model is one component. Section 3 presents the data used in the estimation process. Section 4 contains a description of the statistical model, and Section 5 shows selective results and evaluates the predictive power of the model. Finally, Section 6 discusses policy issues and offers a few conclusions.

2. The Institutional Framework and The Labour Market Reform

Denmark has a two-tiered system for unemployed workers. Most workers in Denmark – around 80% – are members of an unemployment insurance fund. These individuals have, upon the fulfillment of a few conditions, the right to receive unemployment insurance (UI) benefits, which correspond to 90% of the previous wage with an upper limit of approximately 1800 Euro per month. UI benefit payments are heavily subsidized by the state, which finances around 80% of total payments. This system is administered by the Central Labour Market Authority (Arbejdsmarkedsstyrelsen), which is a unit operating under the Ministry of Employment.

Unemployed workers without UI benefit eligibility may instead receive social assistance (SA) benefits. While non-insured workers make up only around 20% of the workforce, they make up a much larger fraction of the unemployed, as the group typically consists of workers with a low attachment to the labour market. Hence, they are more often unemployed, and on average they are unemployed for longer periods. Social assistance benefits are means tested, but typically the amount is below the UI benefit level. Social assistance is administrated by the municipal authorities. There are 279 municipalities in Denmark. Needless to say, they are all subject to the same rules and regulations, but the administration differs considerably between municipalities, and recent research has shown that the differences in efficiency between municipalities in bringing SA recipients back to work cannot be explained solely by individual and municipality-specific variables.[3] In other words, the causes of the differences in efficiency between the local labour market policies are unknown.

3. Arendt et al. (2004).

Up until 2002, the rules and regulations regarding contacts with caseworkers, participation in labour market programs etc. differed between the two systems. The labour market reform of 2002, of which the profiling system is one component, the aims are to eventually have identical rules and regulations in the two systems, and, in fact, to merge the system in the sense that the two-tiered system should become one system. There will still be UI benefits and SA benefits, but the rules regarding meetings, job search etc. will eventually be the same. The goal of the system of reforms is to reduce the emerging public finance problem triggered by an ageing population by increasing the labour force by some 90,000 individuals by 2010. The development of a common profiling model for assessing the employability of unemployed workers marks a step towards a single-tier system.

2.1. Profiling

The profiling model is designed to increase equity and efficiency in the labour market.

If equity is defined in terms of some utility measure, the profiling model will help achieve the goal of increasing equity if it is facilitates helping the unemployed back into work and if the unemployed are in fact among the people in the society who are worst off measured on the utility scale. If instead equity is defined as equal opportunities, the profiling model should be particularly able to increase equity, as it introduces an effective way to provide equal treatment of unemployed workers across different municipalities.

The profiling model will achieve the goal of increasing efficiency if it allocates the labour market programs to the unemployed workers who have the largest (expected) effects of participating in the programs. However, it is not clear that these workers are the same as those who become long-term unemployed, and hence efficiency in this sense also requires a model for targeting active policies. However, efficiency in terms of identifying those at risk of LTU and leaving alone those who are perfectly able to find jobs themselves (thus avoiding deadweight losses) is an important aim of the profiling model; each year a very large number of workers experience a short period of unemployment, and this leads to at least two very important reasons for identifying as quickly as possible those who are at risk of LTU. First, early identification of individuals at risk of LTU allows preventive policies to be implemented during the early stages of unemployment. Second, as mentioned above, early identification is necessary in order to avoid targeting individuals who are perfectly able to find jobs on their own.

The profiling model consists of several components.[4] First, there is a 'job barometer', which is a graphical representation of the predictions based upon the statistical

4. See Arbejdsmarkedsstyrelsen (2004a) for further information about the system.

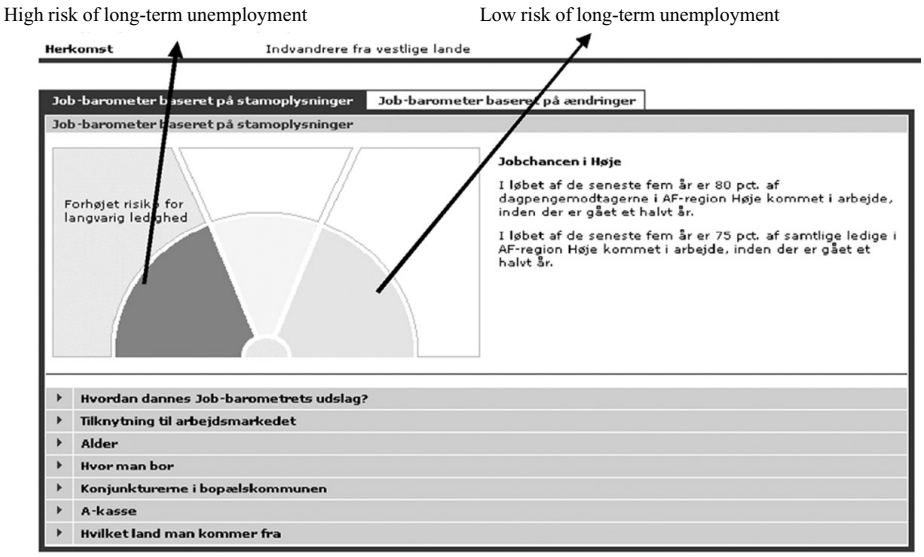

Figure 1. The Job Barometer.

profiling model presented in this paper. This is used by the caseworker to assess employability before the first meeting with a newly unemployed person. Next, there is a public assistance record, which gives the caseworker an overview of the person's previous periods on public assistance.[5] Third, there is a dialogue guide for the caseworkers' communications with clients designed to identify strengths and weaknesses in relation to the labour market. Finally, the unemployed person has to prepare some personal information before the first meeting. This should make it easier across employment offices to treat similar cases alike, and eventually to conduct labour market policies as efficiently as possible.

The aim of the profiling system is to assess the employability of newly unemployed workers. This will be done by eventually placing each individual in one of five categories, ranging from fully employable to (at present) fully unemployable. The statistical profiling tool basically calculates a probability that an individual with certain characteristics – including the labour market history for the past five years – will still be unemployed in six months time, conditional on the elapsed duration of unemployment, which, at the date of the meeting, can be anything from 4 to 30 weeks. This information is presented to the caseworker in the job-barometer, which is shown in Figure 1.

The area to the left indicates 'high risk of LTU', the intermediate area indicates 'medium risk of LTU', and the area to the right indicates 'low risk of LTU'. Which area is highlighted depends on the way the individual's probability deviates from a population mean. The empirical foundation for these probabilities is described in the following sections.

5. This information is also used in the statistical profiling model.

3. The Empirical Model

When faced with the challenge of constructing the statistical tool of a worker profiling model, one must choose an appropriate econometric/statistical model. We first note that the problem of early identification of potentially long-term unemployed workers is related to similar economic problems in the finance literature.[6] Early detection of financially distressed firms with a high risk of bankruptcy, or ways to correctly classify profitable investments have obviously received considerable attention for many years. This has led to the estimation of profiling models by many different techniques including multivariate discriminant analysis, probit and logit models, and more sophisticated nonlinear models such as neural networks models. However, the overall pattern appears to be that the differences between statistical models, in terms of predictive accuracy, is relatively unimportant.[7]

In the U.S., the original Worker Profiling and Reemployment Services (WPRS) model applied a discrete choice model, where the dependent variable was UI benefit exhaustion. Recently, Black et al. (2003) have criticized the WPRS model on that choice. Their main concern is that by using a dichotomous dependent variable all data variations among individuals who do not exhaust their UI benefits are ignored. Instead they suggest in the Kentucky Profiling Model (henceforth KPM), that a continuous dependent variable is employed. Specifically, they suggest, as dependent variable the ratio of benefits drawn to benefit entitlement (i.e., fraction of benefits claimed). In an earlier paper (using the same data and the same dependent variable) several different statistical models are compared, Berger et al. (2000). Among ordinary least squares (OLS), Cox proportional hazard models, and tobit models, the difference in predictive power between the models is very low, and the authors suggest, for simplicity, using the OLS model.

The dependent variable of interest in the Danish context is really an indicator of whether a given individual, conditional on the elapsed unemployment duration, is still unemployed after an additional 6 months. Hence, for each value of the elapsed duration of unemployment, we could create a dichotomous variable taking the value 1 if the individual 'survives' 26 additional weeks in unemployment, and 0 otherwise. We could then estimate a probit or logit model for each value of the elapsed duration of unemployment and use the estimated parameters of those models for predictive purposes. However, such a strategy is also vulnerable to Black et al.'s (2003) criticism, since we do not fully exploit all information in the data to reduce the uncertainty of the parameter estimates. We have therefore chosen, instead, to estimate the duration of

6. Prediction and profiling problems are also seen in various other kinds of literature, such as criminology, Auerhahn (1999), insurance, Yeo et al. (2001), marketing, Shaw et al., (2001) and medicine, Khan et al. (2001).

7. See Altman (1968) for an early discussion and O´Leary (1998) for a more recent discussion of different models.

unemployment and subsequently use the parameters estimated in the duration model to calculate the probability of 'survival' for 26 additional weeks, conditional on the elapsed duration. Hence, our dependent variable of interest is the duration of unemployment, and the econometric/statistical models to be employed are duration models.

Let the continuous stochastic variable T, $T \in (0, \infty)$ denote unemployment duration. The hazard rate, which denotes the probability that an individual with observed characteristics x finds a job in the interval $t + dt$ given that the individual is still unemployed at time t, is then given by

$$h\,(t \mid x_t\,) = \lim_{dt \to 0} \frac{P\,(t < T \leq t + dt \mid T > t,\, x_t\,)}{dt} \tag{1}$$

$$= \frac{f(t \mid \{x_s\}_0^t)}{S(t \mid \{x_s\}_0^t)}, \tag{2}$$

where $f(t \mid \{x_s\}_0^t)$ is the density function, $S(t \mid \{x_s\}_0^t)$ is the survivor function, and $\{x_s\}_0^t$ denotes the entire path of the explanatory variables from the start of the unemployment spell until time t. The survivor function denotes the probability that an individual is unemployed more than t weeks. The association between the hazard function and the survivor function can also be expressed as

$$S(t \mid \{x_s\}_0^t) = \exp\,(-\int_0^t h(s/x_s)ds).$$

The objective of the profiling model is to calculate the probability of remaining in unemployment for more than 26 additional weeks conditional on the elapsed unemployment duration being between 4-30 weeks. Suppressing the dependency on x, this conditional probability can be written as

$$\begin{aligned} Pr(T > \tau + 26 \mid T > \tau) &= \frac{S(\tau + 26)}{S\,(\tau)} \\ &= \frac{\exp(-\int_0^{\tau + 26} h\,(s)\,ds)}{\exp(-\int_0^{\tau} h\,(s)\,ds}, \\ &= \exp\,(-\int_{\tau}^{\tau + 26} h\,(s)ds \end{aligned} \tag{3}$$

where τ denotes the elapsed duration of the unemployment spell. In practice, as mentioned above, $4 < \tau \leq 30$ for individuals in the UI system, since the first interview conducted by the Public Employment Service (PES, henceforth) takes place after 1 month of unemployment. Thus, in the estimations we consider a population that has survived 4 weeks of unemployment. In order to calculate (3) as accurately as possible, we restrict attention to the first 52 weeks of the unemployment spell, that is $T \in [4,56]$. Consequently, all unemployment spells longer than 56 weeks are censored at a duration of 56 weeks. For individuals in the SA system, the first interview may take place from the first day of entry, hence for this system, we will have $0 < \tau \leq 26$, and accordingly we can censor all durations in this system at 52 weeks.

The hazard function is specified as a proportional hazard model. That is, the hazard is the product of the baseline hazard, which captures the time dependence, and a function of observed time-varying characteristics, x_t

$$h\,(t \mid x_t) = \lambda\,(t) \cdot \varphi(x_t), \tag{4}$$

where $\lambda\,(t)$ is the baseline hazard, and $\varphi(x_t)$ is the scaling function specified as $\exp(x_t\beta)$. The baseline hazard is specified as a piecewise constant baseline hazard with splitting times $\tau_0 = 4$, $\tau_1 = 5$, $\tau_2 = 6,...,\tau_{52} = 56$ for individuals in the UI system, that is, there is a separate baseline component for each week. The baseline is defined similarly for the models for the SA system, with $\tau_0 = 0$, $\tau_1 = 1$, $\tau_2 = 2,...,\tau_{52} = 52$. The value of the baseline hazard in the k'th interval is denoted λ_k.

In the scaling function $\exp(x_t\beta)$, the explanatory variables are allowed to be time-varying, as noted above. Let d denote the censoring indicator, which takes the value 1 if the observation is shorter than 56 weeks and uncensored, and zero otherwise.

Let θ denote the parameters of the model. To obtain estimates of the parameters, we perform maximum likelihood estimation based on the following (conditional) log-likelihood function, see Lancaster (1990, for details on duration models).

$$\log l(\theta) = \sum_{i=1}^{N} \left[d_i \ln(h(t_i \mid x_{i,t})) - \int_4^{t_i} h(s_i \mid x_{i,s})\,ds \right],$$

where N denotes sample size. This log-likelihood function is for the models for the UI system. For the SA system, it looks similar, except the lower bound for the integral is 0 instead of 4.

Based on the estimated parameters, the probability that an individual who has been unemployed for τ weeks will experience 26 additional weeks of unemployment is easily calculated as

$$\widehat{Pr}\,(T > \tau + 26 \mid T > \tau, x_\tau) = exp\left(- exp(x_\tau \widehat{\beta}) \sum_{k=\tau}^{\tau+26} \widehat{\lambda}_k \right)$$

assuming that the x does not change.[8]

3.1 Unobserved Heterogeneity

Duration models would typically also include a component intended to capture unobserved heterogeneity. In duration models, it is well known that the baseline hazard is biased towards negative duration dependence if neglected unobserved heterogeneity is present. Moreover, the remaining parameter estimates will be biased too since the model is non-linear.

However, for the present model, the objective is not consistent estimation, but predictive ability. Neglect of unobserved heterogeneity implies that the baseline hazard and the other model parameters will be affected by unobserved heterogeneity. So, for example, if we know that an individual has survived some weeks in unemployment we also know that his or her unobserved characteristics are not that favourable. However, these characteristics are not observed, but their effect is reflected in the baseline hazard. So, for predictive purposes, this seems the best way to exploit all information. If we were to include unobserved heterogeneity, we would be forced to evaluate everyone at the mean (or some other arbitrarily chosen point) of the unobserved variable, and then knowing that the person survived half a year in unemployment is not allowed to influence the evaluation of the hazard.[9] For this reason, the model does not correct for unobserved heterogeneity.

We should also note that the issue of unobserved heterogeneity could influence the degree of success for the profiling model. When evaluating the profiling model and comparing the model to the existing system, where allocation of unemployed workers into labour market programs is, to a large extent, subjectively chosen by the caseworker, we know that the caseworker can observe more detailed information about the individual. Some of the reasons for the unobserved heterogeneity are really data issues and we will discuss these in the section »explanatory variables«, while others may be more interesting; motivation, for example, is not observed by the econometrician but may be partially observed by the caseworker. If it is very important to observe motivation in

8. Since the purpose of the model is to predict whether an individual survives an additional 26 weeks in unemployment, we cannot use time-varying variables in the predictions; the path of the x's is not known in advance. Hence, we make the simplifying assumption that the current value will prevail.

9. Of course, one could also calculate the distribution of unobservables conditional on the elapsed duration of unemployment, from that infer the mean of the unobserved variable given the elapsed duration and use that number for the predictions. Our approach is a shortcut.

order to allocate the unemployed worker into the optimal labour market program, then the caseworker may be able to do a better job than the statistical model. Some experiments from Switzerland indicate, however, that caseworkers have difficulty in predicting treatment effects and so it seems there is indeed room for additional improvements.[10]

4. Data

The analysis here uses data from administrative registers from the Danish Labour Market Authority. This is the same data that the employment offices use and therefore the same information on which predictions have to be made. The advantage of the data set is that it is updated with a very short time lag. The disadvantage is that it basically only contains labour market data. Ideally, we would have liked to use more information by merging to other administrative registers, but since the aim of the analyses is to maximize predictive power based on the available information, we use only what is readily available. The register we use is called DREAM (Danish Register for Evaluation Of Marginalisation), and it is basically an event history file, that includes weekly information on each individual's receipt of public transfer incomes, unemployment registrations, and participation in active labour market programs. Based on the information, a weekly event history is constructed, where the individual each week either occupies one of a number of public transfer states or is not receiving public transfers. When an individual is not registered as receiving public transfers, the person can either be employed or be outside the labour force without receiving transfer income. In the Danish welfare state, the latter is very unlikely; hence the assumption that not receiving public transfers in a given week corresponds to employment is innocuous. From DREAM, we sample the inflow to unemployment in both the UI and the SA systems over the period January 1999 to June 2003. All exits from unemployment to states other than (what we assume to be) employment are treated as independently right censored observations.

For persons in the UI system, we exclude all unemployment durations shorter than four weeks, because the first meeting will never take place during the first four weeks.[11]

Moreover, all unemployment durations longer than 56 weeks in the UI system and longer than 52 weeks in the SA system are censored at these durations, because that is

10. See Lechner & Smith (2005).

11. This truncation from the left also implies that we eliminate a substantial number of temporary layoff spells. Temporary layoff is very common in the Danish labour market since employers only pay UI benefits for the first two days of an unemployment period. Approximately 40 per cent of all unemployment spells in Denmark are temporary layoffs. They are, however, typically quite short and therefore only constitute around 16 per cent of total unemployment, Jensen & Svarer (2003). Moreover, approximately 90 per cent of them are four weeks or shorter.

all the information we are going to use in the estimation process. As the profiling model is further developed, it will eventually be extended such that it can make predictions for persons with an elapsed unemployment duration longer than 26 (or 30) weeks, but to directly extend the current model implies an assumption that the effect of covariates does not change over unemployment duration. Several studies have shown that this assumption is not realistic for longer durations, hence the intention is to estimate new models for elapsed durations over 52 (or 56) weeks, thus essentially allowing for time-varying parameters of the models.

4.1 Sample selection and subsampling

Denmark is divided into 14 counties, plus a 'region' consisting of Copenhagen and Frederiksberg municipalities, each with different labour markets and different local labour market conditions. We follow that division in our estimations below and split the data by region of residence. This leads to 15 sub samples.

As mentioned above, there are two parallel labour market systems, one for workers insured against unemployment (the UI system), and hence eligible for UI benefits, and one for the non-insured (the SA system). Hence, in each region, data are also split according to the labour market system to which each worker belongs.

In each system, different rules apply to different age groups. In the UI system, the Youth Unemployment program applies to workers aged below 25. The data is therefore split into two groups: those under 25, and 25 and older. For workers in the SA system, more active policies are pursued for those aged below 30 than for those above. For workers in the SA system, the data is split into those under 30, and those who are 30 or older. The sample is truncated from below to persons aged 16 or older. In addition, due to a mandatory retirement age of 65, all samples are restricted to those aged 64 or below.

Finally, previous investigations show that male and female workers have very different behaviour in unemployment, so the data is also divided by gender. In total, we thus end up with 15 (regions) $\times$ 2 (systems) $\times$ 2 (age groups) $\times$ 2 (gender) data sets, that is, 120 sub samples of the inflow into unemployment during the period January 1999 – June 2003. The duration model specified above is estimated separately for each of these 120 sub samples.

The dependent variable of the study is the duration of unemployment. In the UI system, the dependent variable is the duration of unemployment given that it is at least four weeks. After these sub sample definitions and the reduction in the samples, we end up with a total of almost 2 million unemployment spells that are used in the estimations.

4.2 Explanatory Variables

Since the purpose of this exercise is to make sound predictions, we use the 'kitchen sink' approach to determin which explanatory variables to include in the model. However, since the data are obtained directly from the Danish Labour Market Authorities, we only have access are variables that are available in their databases. The implication is that the information is usually available when working with Statistics Denmark's register based data is not generally available to us. For example, measures such as education, previous wage, and working experience are not in this data set.[12] The information available is the following:

Age: The individual's age is known, and it is used to construct a set of dummies for age group. In the samples of young individuals (aged below 25 or 30), there is a dummy for each age from 16-29 or 16-24 (with 29 or 24 being the reference category), and for the samples of 'older' individuals, we construct dummies for belonging to 5-year age intervals.

Year: We have included a set of indicators measuring the year in which the unemployment spell begins. As we are only looking at short spells, censoring all spells at 56 weeks, it is not important to take into account time-varying calendar time effects during an unemployment spell.

Municipality: We have a set of indicators – a different set for each county – for the municipality of residence of the unemployed person.

Local unemployment rate: The municipal unemployment rate is included to allow for cyclical effects and thereby improve the predictive power of the model when a new year is entered without the model being updated. This variable is identified because it can vary over time and between municipalities. Hence, it is not perfectly correlated with a linear combination of annual dummies and municipal dummies.

Unmarried: This measures whether an individual is unmarried and does not cohabit either.

Sick: Indicates that an individual is currently reported sick (receiving sick pay) while unemployed. It is thus a time-varying variable.

Immigrant: We have four indicators for whether the individual is first or second generation immigrant from more or less developed countries. The reference category is native Danes.

UI-fund: We have a set of indicators for unemployment insurance fund membership. There are several UI-funds in Denmark, and membership is often categorized according to education/skills and/or by industry. We have 36 different UI funds, and we have an indicator for each. These funds may be seen as broad proxies for the mis-

12. The intention is to increase the information available in the register, so that the caseworker also has this information and so we can base predictions on it. It is not yet available, however.

sing information concerning education and skills. This set of variables is (naturally) only used in the samples of workers insured against unemployment.

Maternity Leave and Holiday Pay: We know whether an individual has been on maternity leave and whether an individual has received holiday pay while unemployed; in employment individuals accumulate rights to holiday payments. If the individual, due to unemployment in the previous year, has not accumulated a sufficient amount of money it is possible to receive money from the state for holidays. Individuals currently employed will count as unemployed in the period they receive vacation pay. We take this into account in the model. These variables are thus time-varying.

Active Labour Market Policies: We have a set of time-varying variables indicating whether the individual is currently in a labour market program, and whether the individual has completed a labour market program during the past 26 weeks. This information is naturally time-varying.

Labour Market History: The most important information, however, for our predictive purposes, is the history of past labour market performance. We know, for each of the five years preceding the current unemployment spell, the fraction of the year spent on some kind of income transfer (UI, SA, temporary leave schemes including parental leave, or other public transfer schemes). We have the same information for sickness periods as well. So we have constructed 10 variables, five for public transfers and five for sickness, measuring the fraction of each of the past five years spent in either sickness or on a transfer scheme. Moreover, we use the number of unemployment spells the individual has had over the same period. For the 'young samples', we only use the information for the past two years. If the information is missing (because the individual was 'too' young), these variables are set at zero.

5. Empirical findings

This section contains a short description of the estimated parameters in the duration model. The main focus of the paper, however, is the predictive power of the model, so this will be brief, and as before mentioned the estimated parameters may be biased due to unobserved heterogeneity. The entire set of estimation results is available on request. To get an idea of the results, table 1 presents the effects of various explanatory variables for insured unemployed men over 25 from Aarhus county, i.e., these are the results for one out of 120 subgroups. Note that we only present a subset of the coefficients, as the UI fund indicators (36), the municipality indicators (279 in total), and indicators for participation in and completion of active labour market programs are not shown.

5.1 Baseline hazard function

In the empirical model we have modelled the baseline hazard as a piecewise con-

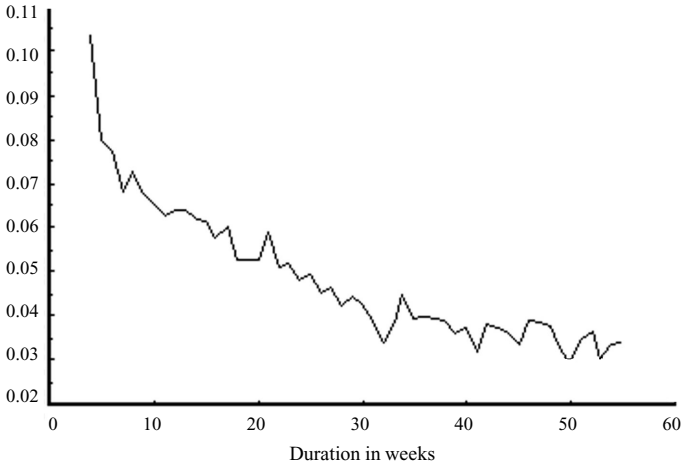

Figure 2. Baseline hazard for Male UI Fund members, over 25, and resident in Aarhus County.

stant function. This very flexible specification is attractive when the baseline hazard exhibits non-monotone behaviour. Watson et al. (1997) impose a Weibull distribution on the baseline hazard.

In figure 2 we show the baseline hazard for insured men over 25 years old in Aarhus county.

We find that the baseline hazard generally exhibits negative duration dependence. This holds for all 120 sub samples. However, it may, to some extent, just reflect neglected unobserved heterogeneity, see the discussion in section 3.1. The peaks in the baseline every 4-5 weeks probably reflect that most jobs begin and end at the start of a month. The hazard rates are generally much lower for persons on social assistance than for persons receiving unemployment insurance benefits.

5.2 Effects of explanatory variables

The effects of some of the explanatory variables differ across the different subgroups, but some consistent patterns arise. The Danish economy has, like most of the western world, experienced an economic downturn after the IT-bubble burst and the September 11 terror attacks in NY. This is captured by the year dummies. They show that compared to the reference year, 1999, the hazard rate out of unemployment has been lower in subsequent years.

As it is witnessed in several studies of unemployment duration, the hazard rate out of unemployment decreases with age. This is also the case in our models. Not surprisingly, men who are out of work due to holiday or paternity leave,[13] leave unemploy-

13. For some reason, such individuals are characterized as unemployed while they are on these transfers.

Table 1. Hazard model for men, insured, above 25 from Aarhus county.

Variables	Coefficients	Std. Error
1999 (reference year)		
2000	*-0.033*	0.016
2001	*-0.036*	0.018
2002	*-0.218*	0.018
2003	*-0.231*	0.029
Age 26-29 (reference age group)		
Age 30-34	-0.025	0.017
Age 35-39	*-0.097*	0.018
Age 40-44	*-0.132*	0.019
Age 45-49	*-0.217*	0.020
Age 50-54	*-0.346*	0.020
Age 55-59	*-0.624*	0.022
Age 60-64	*-0.759*	0.037
Temporarily on Holiday Pay	*1.292*	0.042
Temporarily on Paternity leave	*0.369*	0.114
Temporarily on Sickness benefits	*-0.424*	0.033
Single	*-0.171*	0.011
1. generation immigrant from developed country	*-0.176*	0.029
1. generation immigrant from less developed country	*-0.356*	0.030
2. generation immigrant from developed country	-0.095	0.109
2. generation immigrant from less developed country	*-0.401*	0.134
Sickness benefit rate 1 year ago	*-0.289*	0.104
Sickness benefit rate 2 years ago	0.022	0.225
Sickness benefit rate 3 years ago	-0.013	0.353
Sickness benefit rate 4 years ago	-0.140	0.359
Sickness benefit rate 5 years ago	-0.197	0.241
Public transfers rate 1 year ago	*-0.088*	0.043
Public transfers rate 2 years ago	-0.159	0.092
Public transfers rate 3 years ago	0.019	0.137
Public transfers rate 4 years ago	0.017	0.168
Public transfers rate 5 years ago	*-0.686*	0.106
Number of unemployment spells last year	*0.054*	0.010
Number of unemployment spells two years ago	*0.087*	0.006
Local unemployment rate	*-0.503*	0.135
36 UI Fund Membership Indicators	Yes	
26 Municipality of Residence Indicators	Yes	
ALMP Participation and Completion Indicators	Yes	

Note: Italic figures indicate that the parameter is different from 0 at the 5% level. In the regression we also corrected for municipality effects for UI-fund membership, and for participation in ALMPs.

ment faster. In addition single men are less likely to leave unemployment compared to their married or cohabiting counterparts. This result is consistent with previous investigations of unemployment duration. Being a first or second generation immigrant

from less developed countries is associated with lower hazard rates and therefore longer unemployment durations. For immigrants from developed countries, the same pattern emerges, but it is less clear and the coefficients are smaller and more often insignificant. Second, the larger the fraction of time in the past five years spent on transfer incomes, the lower the probability of leaving unemployment. The same results hold for sickness periods. However, when we look at the number of spells, we find that the more unemployment spells an individual has had, e.g. during the past two years, the higher is the hazard rate out of unemployment. This coefficient, however, must be interpreted *given* the level of the variables reflecting the fraction of time spent on transfer schemes. That is, individuals with many short spells of unemployment in the past are also likely to have a short current spell of unemployment. Finally, we see that the local unemployment rate influences the hazard rate out of unemployment. This is consistent with e.g. Svarer et al. (2004). They find that the mobility among the unemployed is very low in Denmark. As a consequence, people tend to be unemployed longer if they stay in a region with a high unemployment rate.

5.3 Assessment of predictive power

The primary purpose of this exercise is to construct a tool that can guide case-workers in their work. The prime success criterion is of course that they can trust the outcome of the statistical model. Consequently, we are interested in identifying the group of newly registered unemployed that has the highest probability of experiencing more than 26 weeks of unemployment. We will denote that group 'potentially long-term unemployed' (PLTU) whereas their counterparts are the potentially short-term unemployed (PSTU). Define as the *cut-off value* the number which $\widehat{Pr}\,(T > \tau + 26 \mid T > \tau, x_\tau)$ shall exceed in order for an individual to be identified as PLTU. We can subsequently calculate the number of correct predictions; that is, the number of actually short-term unemployed workers (those with unemployment spells shorter than 26 weeks) who are also predicted to be short-term unemployed, plus the number of actual long-term unemployed workers who are also predicted to be long-term unemployed. This number can be related to the number of incorrect predictions. The choice of *cut-off* value will clearly have an effect on the outcome of this comparison. We have chosen to determine the *cut-off* value (separately for each of the 120 sub groups) so that the following sum is maximized:[14]

Number of short-term unemployed predicted to be short-term unemployed +
Number of long-term unemployed predicted to be long-term unemployed

14. This objective functions implies that the two groups are weighted by their relative sizes. By simply changing the weights it is possible to put more emphasis on one group if desired.

Table 2. The distribution of predictions and actual outcomes.

Groups	Fraction of correct predictions	STU who are PSTU*	STU who are PLTU	LTU who are PSTU	LTU who are PLTU	No. of observations
Women, ≤ 24, ins.	0.68	29,879	1,822	13,983	3,283	48,967
Women, ≤ 29, not ins.	0.66	71,485	18,182	34,466	29,709	153,842
Women, ≥ 25, ins.	0.61	174,840	70,063	110,001	104,331	459,235
Women, ≥ 30, not ins.	0.68	16,371	18,845	7,849	39,458	82,523
Men, ≤ 24, ins.	0.80	46,616	452	11,188	580	58,836
Men, ≤ 29, not ins.	0.68	102,444	12,030	43,316	16,004	173,794
Men, ≥ 25, ins.	0.71	271,420	25,947	104,353	36,474	438,194
Men, ≥ 30, not ins.	0.64	39,558	23,950	21,114	41,114	125,736
Total	0.66	752,613	171,291	346,270	270,953	1,541,127
Percentage		0.49	0.11	0.22	0.18	

* PSTU: Predicted to be short-term unemployed. PLTU: Predicted to be long-term unemployed.

Note that when making an assessment of the correct predictions, we are forced to leave out all spells that are right censored at a duration shorter than 30 weeks in total. Moreover, when making predictions, we use only the value of explanatory variables at the beginning of the spell. That implies that we could make even better predictions if time-varying variables were taken into account in the process of prediction. Table 1 contains the aggregate numbers of correct and incorrect predictions for the entire country.

The fraction of correct predictions is 0.66. Compared to the New Zealand profiling model that also employs a duration model, we gain a significant improvement in predictive power. In their model they are able to make 59 per cent correct predictions. Since these fractions are dependent on the populations the models are trying to categorize, in particular the relative size of the groups, it is probably more interesting to consider the fraction of correct predictions in each group. 82% of the STU are predicted correctly while 55% of the LTU are predicted correctly, and the corresponding fractions are both equal to 59% in the New Zealand model. The key to the improved predictions is obviously the sub sampling and the large number of variables, especially the information on past labour market history, which greatly improves the predictive power of the model.[15] Furthermore a different objective function is used in their study, which in effect amounts to assigning equal weights to the two categories.

Looking across subgroups, it is revealed that the predictive power is higher for men than for women and for younger workers compared to their older counterparts. The former result is in line with previous research on modelling of individual unemployment.

15. The variables used in the New Zealand model are: age, prior unemployment, gender, ethnicity, qualifications, urban location and regional location.

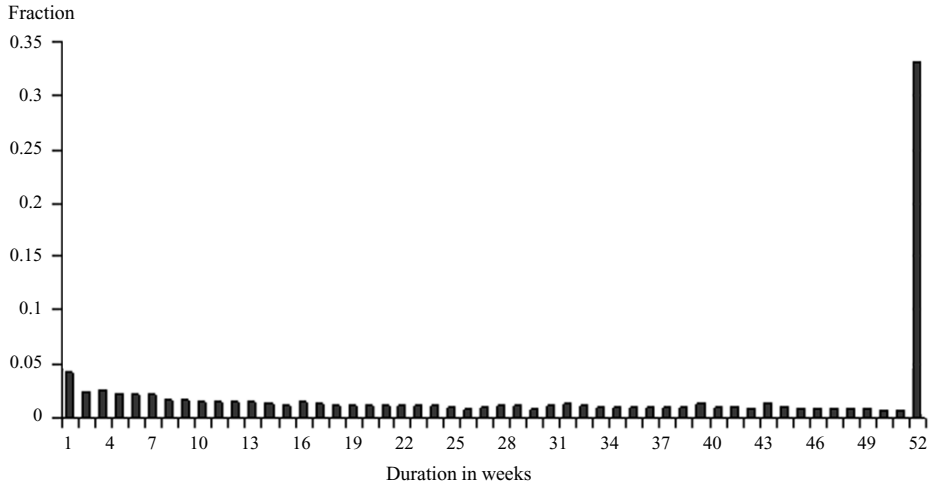

Figure 3. Histogram for actual duration of those who are PLTU.

We also performed out-of-sample predictions. In practice we randomly divided all the samples in halves. We estimated the model on the first half of the data and applied the parameters' estimates to the second half for predictions. The predictions were almost identical to their full-sample counterparts.

Even though we obtain a reasonable level of correct predictions, there are still a substantial number of individuals who – if the model's predictions were taken at face value – would be put into the wrong category, but that is exactly the reason why the statistical model is only a part of the profiling system. As discussed above, it is an input the caseworker can use to extract useful information regarding the potential risks of LTU facing an unemployed worker. In future versions, the profiling model will include information about unemployed workers gathered by the caseworkers. This information will give an impression of how motivated the individual is in terms of regaining employment, how employable the person is etc. When this information becomes available, we expect to have a profiling model that is even better at predicting the LTU risk.

As a result of the estimation procedure with right censored observations it is not very informative to compute expected values and standard deviations. Instead we turn to a graphical depiction of some of the issues concerning the precision of the predictions. In figure 3 we look at the sub sample consisting of unemployed males in Aarhus who are insured and more than 25 years old, and see that the classification into the two categories is not nearly perfect. A considerable number of the PLTU find a job before 26 weeks, though the majority are unemployed for a longer period and 33%, actually, than a year. Similarly some of the PSTU are unemployed for more than 26 weeks but still, the majority get a job earlier, see Figure 4.

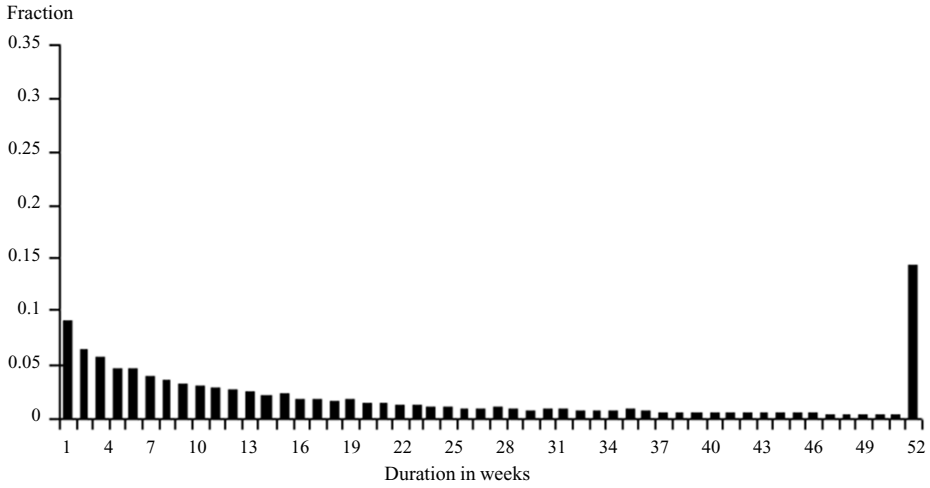

Figure 4. Histogram for actual duration of those who are PSTU.

Another way to illustrate the prediction ability is to look at the fractions that are predicted correctly distributed over the actual experienced unemployment duration. Figure 5 shows this for the particular sub sample.

It should be noted that the reason for the low fractions of PLTU is that in this particular sub sample the group of STU is much bigger than the group of LTU and hence, predicting STU correctly receives a relatively high weight in the objective function, which is to maximize the number of correct predictions. However, abstracting from the low level, it is clearly seen that the fraction of correctly predicted long-term unemployed is increasing in the duration, so that we are more likely to correctly classify the »really long-term« unemployed, i.e. those who experiences more than a year of unemployment. This should then be compared to the cost of making incorrect predictions for different types of unemployed workers. If it is most costly to classify »really short-term« unemployed as PLTU and »really long-term« unemployed as PSTU compared to making mistakes in the interval in between, then we would like to have the highest fractions of correct predictions at very short and very long durations. But to the extent that the profiling system should work as an additional tool for the caseworker, it might actually be more valuable to be able to distinguish those in between if the caseworkers themselves are capable of correctly identifying the extremes. Another way to put this is that it could be a partial degree of explanation rather than an absolute that we want to maximize. In the discussion of this issue, it is also important to remember that we are trying to make a rather crude categorization of the unemployed into two groups and the results should hence be interpreted with this in mind.

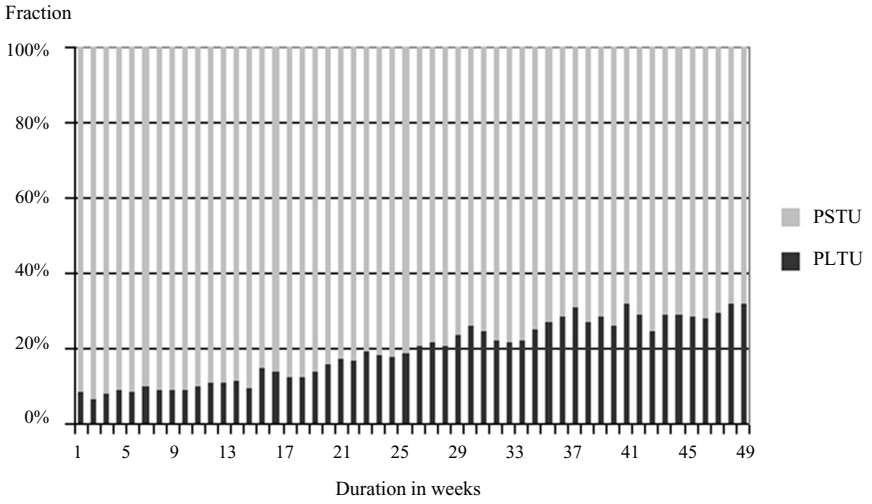

Figure 5. Fractions of PLTU and PSTU over actual duration.

6. Concluding remarks

In this paper we describe a statistical model used for profiling of newly unemployed workers in Denmark. When a worker – during his or her first six months in unemployment – enters the employment office for the first time, this model predicts whether he or she will be unemployed for more than six months from the current date or not. The caseworker's assessment of how to treat the person is partially based upon this prediction. The model – which performs relatively well in terms of predicting actual unemployment – is the first step in the process of developing statistical procedures to assist caseworkers in Denmark in their effort to bring unemployed individuals back into employment. Future amendments to the model include additional information based on caseworkers' assessments of the unemployed individuals in order to make dramatic improvements to the model's predictive ability. Moreover, assessment of the effects of participation in various active labour market programs, that is, a targeting system, seems to be the natural next generation in the world of profiling models.

Literature

Altman, E. I. 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *The Journal of Finance*, Vol. 23, No. 4, 589-609.

Arbejdsmarkedsstyrelsen. 2004a. Visitationsværktøjskassen Version 3, *http://www.ams.dk/visitationsprojekt/visitationsværktøjskassen/v3/Samlet_visitation sværktøjskasse_v3.pdf*

Arbejdsmarkedsstyrelsen. 2004b. En ny måde at møde ledige på, *http://www.ams.dk/visitationsprojekt/Visitation_december_2004.pdf*

Arendt, J. N., E. Heinesen, L. Husted, B. Colding & S. H. Andersen. 2004. *Kontanthjælpsforløbs varighed og afslutning: Forskelle mellem kommuner*. Akf rapport, September 2004.

Auerhahn, K. 1999. Selective Incapacitation and the Problem of Prediction, *Criminology*, Vol. 37, 4, 703-34.

Berger, M. C., D. A. Black & J. Smith. 2000. Evaluating Profiling as a Means of Allocating Government Services, In Michael Lechner and Friedhelm Pfeiffer (eds.), *Econometric Evaluation of Active Labour Market Policies*, Heidelberg: Physica,: 59-84.

Black, D. A., M. Plesca, J. A. Smith & S. Shannon. 2003. *Profiling UI Claimants to Allocate Reemployment Services: Evidence and Recommendations for States*, Final Report to United States Department of Labour.

Frölich, M., M. Lechner & H. Steiger. 2003. Statistically Assisted Programme Selection – International Experiences and Potential Benefits for Switzerland, Swiss *Journal of Economics and Statistics*, 139, 311-31.

Jensen, P. & M. Svarer. 2003. Short- and Long-Term Unemployment: How do Temporary Layoffs Affect this Distinction?, *Empirical Economics* 28, 1, 23-44.

Khan, J., J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson & P. S. Meltzer. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicin*, Vol. 7, 6, 673-79.

Lancaster, T. 1990. *The Econometric Analysis of Transition Data*, Cambridge University Press, Cambridge.

Lechner, Michael & Jeffrey Smith. 2006. What is the Value Added by Caseworkers?, Forthcoming in *Labour Economics*.

OECD. 1998. *Early Identification of Jobseekers at Risk of Long-Term Unemployment*, Paris.

O'Leary, D. E. 1998. Using Neural Networks to Predict Corporate Failure, *International Journal of Intelligent Systems in Accounting, Finance & Management*, Vol. 7, 187-97.

Shaw, M. J., C. Subramaniam, G. W. Tan & M. E. Welge. 2001. Knowledge management and data mining for marketing, *Decision Support Systems* 31, 127-37.

Svarer, M., M. Rosholm & J. Munch. 2005. Rent Control and Unemployment Duration, *Journal of Public Economics.* Vol. 89, 11-12, 2165-81.

Watson, R., D. Maré & P. Gardiner. 1997. Predicting the Duration of Unemployment Spells, *Labour Market Bulletin*, 1997:2, 51-65.

Yeo, A. C., K. A. Smith, R. J. Willis & M. Brooks. 2001. Clustering Technique for Risk Classification and Prediction of Claim Costs in the Automobile Insurance Industry, *International Journal of Intelligent Systems in Accounting, Finance & Management*, Vol. 10, 39-50.