AUTHORS Markus Viljanen, Tapio Pahikkala

TITLE Predicting Unemployment with Machine Learning Based on Registry Data

YEAR 2020

DOI 10.1007/978-3-030-50316-1_21

VERSION Final draft

CITATION Viljanen M., Pahikkala T. (2020) Predicting Unemployment with Machine Learning Based on Registry Data. In: Dalpiaz F., Zdravkovic J., Loucopoulos P. (eds) Research Challenges in Information Science. RCIS 2020. Lecture Notes in Business Information Processing, vol 385. Springer, Cham. https://doi.org/10.1007/978-3-030-50316-1_21

Predicting Unemployment with Machine Learning Based on Registry Data

Markus Viljanen and Tapio Pahikkala

Turun Yliopisto, Turku, Finland
majuvi@utu.fi

Abstract. Many statistical models have been developed to understand the causes of unemployment, but predicting unemployment has received less attention. In this study, we develop a model to predict the labour market state of a person based on machine learning trained with a large administrative unemployment registry. The model specifies individuals as Markov chains with person specific transition rates. We evaluate the model on three tasks, where the goal is to predict who has the highest risk of escaping unemployment, becoming unemployed, and being unemployed at any given time. We obtain good performance (AUC: 0.80) for the machine learning model of lifetime unemployment, and very good performance (AUC: 0.90+) to the near future when we know the recent labour market state of a person. We find that person information affects the predictions in an intuitive way, but there still are significant differences that can be learned by utilizing labour market histories.

Keywords: Unemployment · Machine Learning · Prediction.

1 Introduction

Understanding and predicting unemployment is very important for societies and individuals alike. Identifying at risk individuals and the total amount of unemployment is a central topic of interest, regardless of the context. To understand how unemployment is experienced at the individual level, we need to consider the full labor market history of each person. These histories consists of recurrent spells of unemployment and other labor market states, which together determine the total time person spends in unemployment. When we have a well-founded model for the labor market histories, we can predict these the transitions in and out of unemployment, including the resulting lifetime unemployment.

In this study, we develop a model for the sequence of labour market states of a person. We focus on prediction and evaluate models on three related prediction tasks: what is the probability that a person exits unemployment, becomes unemployed, and is unemployed at any given time? To do this, we model the unemployment status as a Markov chain with person specific transition rates. The prediction of a person's unemployment status is then given by the state probabilities of the Markov chain. The steady state probabilities imply that, in the

long run, an individual can be predicted to spend a certain amount of their lifetime on unemployment. We investigate both a statistical model and a machine learning model for the transition rates in different settings, where predictions can be required for a future time or a completely new person. As a result, we obtain predictive models for unemployment dynamics at the individual level, using simple models trained with few years of historical data.

2 Related Work

Macroeconometric research has studied unemployment trends using historical time series and aggregate labor market statistics. See for example a review [1], which summarized the current state of research on unemployment dynamics in macroeconomic models. It is possible to explain the unemployment rate in terms of the unemployment entry and exit flows, which often correlate with changes in the economy [2]. However, it has been pointed out that individuals have significant heterogeneity, which this analysis ignores [3]. In our Finnish study region of Varsinais-Suomi, the ELY-center (Centre for Economic Development, Transport and the Environment) is required by law to publish the total number of job seekers and basic statistics related to them each month. Recently, KELA (The Social Insurance Institution of Finland) published a working paper which used macroeconometric data to produce estimates on the total amount of lifetime unemployment in the population [4]. However, macroeconometric statistics cannot be used for prediction at the individual level.

Microeconometric studies have been used to develop models at the individual level. In these studies, regression is often used to assess the effect of a variable of interest, such as a policy change in unemployment benefits. The most common application of person-level data is the study of unemployment duration. For example, an older review [5] summarized many studies on how health, age, gender, unemployment benefits, etc. affect unemployment. A recent study [6] reviewed the accumulated evidence of the individual experience of unemployment, predictors of exiting unemployment, and the effect of interventions. In Finland, studies have reported how individual characteristics influence the risk of exiting unemployment [7–9] and the risk of becoming unemployed [10, 11]. Studies have also investigated the flows of individuals between different labor market states, where Finnish studies have been performed on both aggregate [12] and person-level data [13].

Modeling unemployment with Markov Chain models, which is an idea underlying our model, has been considered in a statistical context [14–16]. These studies have focused on the dynamics of unemployment and the effect of different variables, not on the predictive ability of the model. In the predictive task, several studies have investigated the predictive power of Google searches in forecasting the unemployment rate, see for example the references in [17]. In the Finnish context, work at ETLA (The Research Institute of the Finnish Economy) has also predicted the unemployment rate using Google searches [18]. In addition to standard time-series models, machine learning can be applied

to macroeconomic time-series prediction [19]. At the individual level, multiple studies have predicted the Long-Term Unemployment (LTU) status of a person, see for example the references in [20]. This is a simpler task of binary prediction whether an individual would fall under the classification 'long-term unemployed'. The unemployment exit rate was recently studied in [21]. Studies have traditionally used logistic regression, but machine learning algorithms such as gradient boosting or random forest have been found to perform slightly better.

3 Data set

Our data set is based on individual-level administrative data collected in the Varsinais-Suomi ELY-centre, based in turn on the URA-registry collected in the local employment and business services office (TE services). Unlike many other unemployment studies, we do not have a separate questionnaire or other data sets merged into this data set. The registry contains all job seekers registered at the unemployment agencies, which they are required to do in order to receive unemployment benefits. For this reason, practically all unemployed persons can be considered to belong to the registry. For example, the official unemployment statistics produced by the Ministry of Labour (TEM) are based on the number of people in these registries. The administrative registry is not biased by sample selection or subjective reporting. However, if we wish to apply the model to predict unemployment for the Varsinais-Suomi population, the training set should strictly be a random sample of the population. In the appendix, we investigate how the unemployment registry differs from this population.

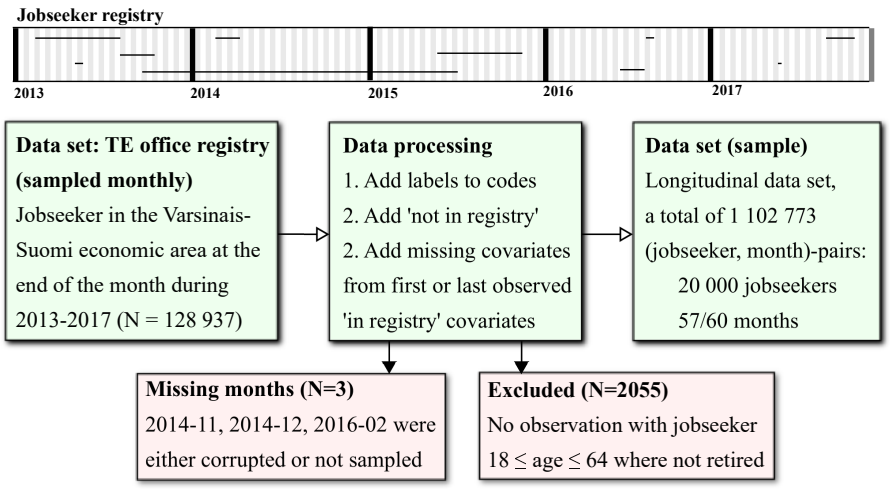

Fig. 1. Process of transforming the unemployment registry to the study data set.

The registry was sampled at the weekday following the end of each month, from the beginning of 2013 to the end of 2017, as illustrated in 1. This resulted in one sample per month over 5 years, in principle 60 monthly samples. The data acquired for this research turned out to have some missed samples due missed manual collection or data corruption, with 57 total monthly samples. The models we use are designed to handle censoring, so this does not bias the results. Each entry identifies the individual by their social security number, records their labor market state and collects some background information. The social security number was replaced by an ordinal userid before analysis due to data protection laws. The original data set had 128 937 potential persons, of which the study used a random sample of 20 000.

The labor market state of a job seeker is recorded as employed on placement, employed, unemployed, laid off, shortened workweek, outside the labour force, and unemployment retirement. If a person is employed or outside the labor market they may be missing from the registry, because they do not have to report to the unemployment office. We have therefore denoted the state of missing from the data with a new code 'not in registry' and filled these observations with covariates primarily forward and secondarily backward from the last observed covariates. We also created a new category of 'censored' observations. These included samples which were either not collected or where the person's age would have been smaller than 18 or over 64.

The individual registry entry has information such as gender, age, work experience, level of education, field of education, field of profession, mother tongue, and citizenship. This information varies over time and these are the possible time-varying covariates to include in the model. However, some of the information has a high degree of identifiability, so we had to use the following higher level covariates: gender, work experience, age in 5 year buckets, level of education, field of education. The background information is recorded in the registry as codes displayed in 1, so we attached human readable labels from Statistics of Finland descriptions. If a code was missing or the corresponding label did not exist, we replaced the covariate value by the 'unknown' category.

Table 1. The following covariates were included in the model.

Registry Entry	Covariate (categories)	Example values
henkilotunnus	userid (20000)	1,2,3...
supukoodi	gender (2)	M, F
ika	age (10)	[18,20), [20,25), ..., [60,65)
tyokokemuskoodi	work experience (3)	None, Some, Sufficient
koulutuskoodi	level of education (16)	Early Education, Basic 1-6, ..., PhD
	field of education (12)	Preparatory, Education, ..., Services
vvvvkk	time (59)	2013-01, 2013-02, ..., 2017-12
voimolevatyollkoodi	labour market status (7)	Employed, Unemployed, ..., Censored

It is well-known that unemployment is influenced by many economic factors: economic growth, labour market conditions, regulation, unemployment benefits, etc. The unemployment rate also fluctuates significantly with the economic cycle. It would be possible to include macroeconomic indicators. However, if one wants to predict future unemployment instead of historical unemployment, one needs to know these variables in the future. This is of course not possible, and predicting them accurately is very difficult. Our data consists of a complete economic cycle in the local unemployment rate (Appendix Figure 6), which implies that the future predictions are time-averaged. This suits us well since we are interested in the long run individual unemployment and do not wish to model the economic cycle. In addition, our evaluation metric depends only on the relative unemployment; it should not affect the performance estimates in any case.

4 Model

4.1 Mathematical framework

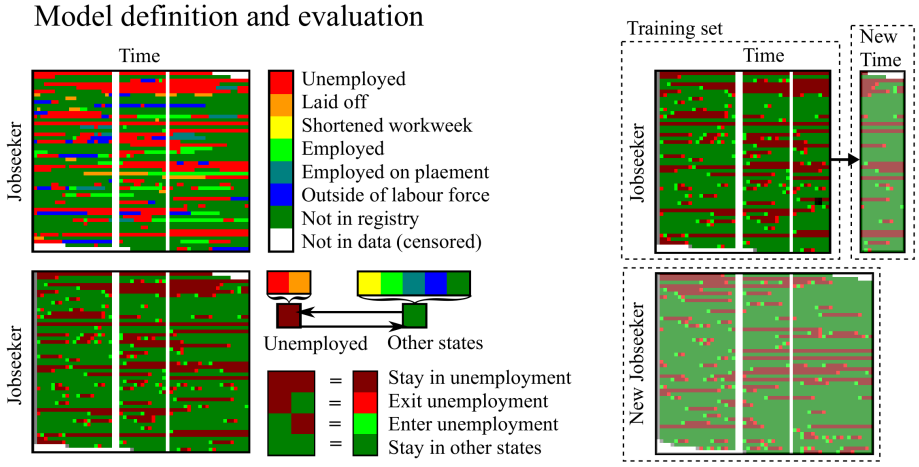

Fig. 2. Model definition and evaluation from the sequence of labor market states.

The model is motivated by the visualization of labor market histories in Fig. 2. For the purpose of this study, we define the labour market status as either unemployed or not unemployed. We treat persons that were either unemployed or laid off as 'unemployed' and other labor market states as 'not unemployed'. Each labour market history therefore is a binary vector of 'unemployed' states $x_i(t) \in \{0, 1\}$ of the individual $i = 1, \ldots, N$ at months $t = 1, \ldots, T$. We describe the labour market history of each person is a stochastic process $\{X_i(t), t \geq 0\}$ where the random variable $X_i(t) \in \{0, 1\}$ is the unemployment status of the

individual i at time t. The random vector $P_i(t) = (P(X_i(t) = 0), P(X_i(t) = 1))$ is the state probability vector, giving the probability of being either in or out of unemployment at a given time. At the prediction time, we know the current state of a person and wish to predict the future state of the person. For this reason, we define the transition probability matrix:

$$\mathbb{P}_i(s,t) = \begin{pmatrix} P(X_i(t)=0|X_i(s)=0) & P(X_i(t)=1|X_i(s)=0) \\ P(X_i(t)=0|X_i(s)=1) & P(X_i(t)=1|X_i(s)=1) \end{pmatrix} \tag{1}$$

The labor market history of a person does not appear to consist of randomly dispersed unemployment events, but rather consecutive months of unemployment which can be characterized as unemployment spells. In addition, some persons seem to have considerably longer unemployment spells than others. This observation is the basis of our model. Every calendar month, a person either remains in the current state or transitions into unemployment (entry) or out of unemployment (exit). We consider that only one transition occurs within each calendar month and that the transition probabilities are person specific. In other words, we hypothesize that the data set is a Markov Chain with person specific transition rates.

If the labor market states at $X_i(t)$ and $X_i(t+dt)$ are different, this means that either of two transition events occurred in $(t, t+dt]$: exit from unemployment or entry into unemployment. Transition rates are defined through the instantaneous probability of a transition. Define the person specific rate of unemployment exit $\lambda_i(t)$ and the rate of unemployment entry $\mu_i(t)$:

$$\begin{aligned} \lambda_i(t) &= \lim_{dt\to 0} P(X_i(t+dt)=0|X_i(t)=1)/dt \\ \mu_i(t) &= \lim_{dt\to 0} P(X_i(t+dt)=1|X_i(t)=0)/dt \end{aligned} \tag{2}$$

The Markov Chain is then defined by a rate matrix $d\mathbb{A}_i(t) = \begin{pmatrix} -\mu_i(t) & \mu_i(t) \\ \lambda_i(t) & -\lambda_i(t) \end{pmatrix} dt$. It can be shown that the transition probability matrix $\mathbb{P}_i(s,t)$ can be computed as the following matrix-valued product integral [22]:

$$\mathbb{P}_i(s,t) = \prod_{u\in(s,t]} (\mathbb{I} + d\mathbb{A}_i(u)) \tag{3}$$

We train the model using time-varying transition rates. However, we predict to the future using last known covariates so that the predicted future rates are constant. Under the assumption of constant rates, we can derive the following interesting results for person's lifetime unemployment using well-known properties of the Markov Chain. The lifetime unemployment is defined as the proportion of the working age life that a person spends in unemployment. Denote the rate of unemployment exit by λ_i and the rate of unemployment entry by μ_i. At prediction time, we wish to predict the future state $\mathbb{P}_i(t+\delta t, t)$ given δt months forward from the last known state $X_i(t)$. In this case, we can explicitly derive a closed form expression for the transition probability matrix to obtain [18]:

$$\mathbb{P}_i(t, t+\delta t) = \begin{pmatrix} \frac{\lambda_i}{\lambda_i+\mu_i} + \frac{\mu_i}{\lambda_i+\mu_i} e^{-(\lambda_i+\mu_i)\delta t} & \frac{\mu_i}{\lambda_i+\mu_i} - \frac{\mu_i}{\lambda_i+\mu_i} e^{-(\lambda_i+\mu_i)\delta t} \\ \frac{\lambda_i}{\lambda_i+\mu_i} - \frac{\lambda_i}{\lambda_i+\mu_i} e^{-(\lambda_i+\mu_i)\delta t} & \frac{\mu_i}{\lambda_i+\mu_i} + \frac{\lambda}{\lambda_i+\mu_i} e^{-(\lambda_i+\mu_i)\delta t} \end{pmatrix} \quad (4)$$

Regardless whether the person is unemployed at the last observation, the transition probability matrix implies that the probability of being unemployed converges over time to:

$$\lim_{\delta t \longrightarrow \infty} P_i(t+\delta t) = \left(\frac{\lambda_i}{\lambda_i+\mu_i}, \frac{\mu_i}{\lambda_i+\mu_i} \right) \quad (5)$$

In the long run, the person with rates λ_i and μ_i can be expected to be unemployed $\frac{\mu_i}{\lambda_i+\mu_i}$ of the time. We call this the individual's unemployment prevalence implied by the transition rates. This observation also means that we lose predictive information the longer we predict to the future.

4.2 Training models

We train the models using the probabilities implied by the mathematical framework. The framework is defined in continuous time, but our data is recorded using one month granularity and we assume that there can be only one transition within a given month. If different months are coded as integers $t = 0, 1, \ldots$ and $\lambda_i(t)$ or $\mu_i(t)$ define the transition rates within that month, we have the probability of a transition [23]:

$$\begin{aligned} P(X_i(t) = 1 | X_i(t-1) = 0) &= 1 - \exp(-\lambda_i(t)) \\ P(X_i(t) = 0 | X_i(t-1) = 1) &= 1 - \exp(-\mu_i(t)) \end{aligned} \quad (6)$$

To fit the model, we think of each labour market history as a sequence of transitions between unemployment and employment as demonstrated in Fig. 2. Denote by $c_i(t) \in \{0, 1\}$ the observability status of a transition, with $c_i(t) = 1$ indicating that months t and $t-1$ were observable for person i and $c_i(t) = 0$ when either was censored. The likelihood of the sequence can be split into two parts. First, the part when the previous observation month t is unemployed where $\mathbb{N}_{i,1} = \{t \in \mathbb{N} : x_i(t-1) = 1, c_i(t) = 1\}$:

$$L_{\lambda_i}(x_i(t)) = \prod_{t \in \mathbb{N}_{i,1}} (1 - \exp(-\lambda_i(t)))^{\mathbb{I}(x_i(t)=0)} \exp(-\lambda_i(t))^{\mathbb{I}(x_i(t)=1)} \quad (7)$$

Second, the part when the previous observation month t is not unemployed where $\mathbb{N}_{i,0} = \{t \in \mathbb{N} : x_i(t-1) = 0, c_i(t) = 1\}$:

$$L_{\mu_i}(x_i(t)) = \prod_{t \in \mathbb{N}_{i,0}} (1 - \exp(-\mu_i(t)))^{\mathbb{I}(x_i(t)=1)} \exp(-\mu_i(t))^{\mathbb{I}(x_i(t)=0)} \quad (8)$$

The likelihood of a person's labour market history is then:

$$L_{\lambda_i,\mu_i}(x_i(t)) = L_{\lambda_i}(x_i(t)) L_{\mu_i}(x_i(t)) \quad (9)$$

We still have to model the person specific transition rates $\lambda_i(t)$ and $\mu_i(t)$. We assume that the transitions are determined by observed characteristics (covariates), unobserved characteristics and randomness in finding or exiting a job. The observed characteristics influence the transition rates through a time-varying covariate vector $z_i(t)$ and a parameter vector α or β, depending on the transition. The unobserved characteristics are modelled by person-specific intercept u_i or v_i, depending on the transition. For example, there could differences in the demand for the occupations, motivations in finding a job, personal issues, etc. We assume that the transition rates follow a proportional rates assumption. This is equivalent to the proportional hazards model in a survival analysis of a single spell with a subject-specific frailty term [16]. The model then defines the person specific transition rates:

$$\begin{aligned} \lambda_i(t) &= \exp(\alpha^T z_i(t) + u_i) \\ \mu_i(t) &= \exp(\beta^T z_i(t) + v_i) \end{aligned} \tag{10}$$

Interestingly, the prevalence $P(X_i(t) = 1) = \frac{\mu_i}{\lambda_i + \mu_i}$ and time-invariant covariates z_i imply a logistic regression model of lifetime unemployment prevalence:

$$\frac{P(X_i(t) = 1)}{1 - P(X_i(t) = 1)} = \frac{\mu_i}{\lambda_i} = \exp((\beta - \alpha)^T z_i + (v_i - u_i)) \tag{11}$$

In a statistical model, the covariates are known as fixed effects and the person specific intercepts are random effects. The model is therefore a two-state mixed effects model. To model the random effects, we assume that they follow a multivariate normal distribution $(u_i, v_i) \sim \text{Normal}(\gamma, \Sigma)$ with a 2×1 mean vector γ and a 2×2 covariance matrix Σ as unknown parameters. This allows a correlation between the unemployment entry and exit rates, since it is possible that persons who have a difficult time of finding employment might also find it difficult to remain employed. Given a data set $D = \{x_i(t)\}_{i=1,\ldots,N}$, we define the unconditional data likelihood by integrating out the unknown random effects using the normal distribution density function $f_{\gamma,\Sigma}(u_i, v_i)$ [25]:

$$L_{\alpha,\beta,\gamma,\Sigma}(D) = \prod_{i=1,\ldots,N} \int_{u_i,v_i} L_{\lambda_i,\mu_i}(x_i(t)) f_{\gamma,\Sigma}(u_i, v_i) du_i dv_i \tag{12}$$

The model is then fit by minimizing the negative log likelihood:

$$\text{argmin}_{\alpha,\beta,\gamma,\Sigma}[-\log(L_{\alpha,\beta,\gamma,\Sigma}(D))] \tag{13}$$

In a machine learning model, we assume that the person specific intercepts u_i or v_i are the elements of a model parameter vector u or v, just like α or β, and make the problem well-conditioned by regularization. This means that we fit the maximum likelihood with a penalty term which is multiplied by a constant C. We set the optimal constant with 10-fold cross-validation by splitting the training set into train and validation sets. However, as long as the solution is defined for $C > 0$, the evaluation was not sensitive to the choice of regularization and we report results for the default value of $C = 1$. We define the conditional data likelihood by assuming that the person specific rates are model parameters:

$$L_{\alpha,\beta,u,v}(D) = \prod_{i=1,\ldots,N} L_{\lambda_i,\mu_i}(x_i(t)) \tag{14}$$

The model is then fit by minimizing the penalized negative log likelihood:

$$\text{argmin}_{\alpha,\beta,u,v}[-\log(L_{\alpha,\beta,u,v}(D)) + C(\|\alpha\|^2 + \|\beta\|^2 + \|u\|^2 + \|v\|^2)] \tag{15}$$

The statistical model and the machine learning model result in surprisingly similar estimators. The person specific intercepts u_i and v_i, which make application of straightforward regression ill-conditioned, are 'shrunk' toward the population averages. The machine learning model is considerably faster to train and yields almost equivalent predictions, though it is not based on a statistical analysis of the problem in question.

5 Results

5.1 Prediction tasks

There are three natural prediction tasks that our model answers:

1. Exit: Who has the highest risk of exiting unemployment?
2. Entry: Who has the highest risk of entering unemployment?
3. Prevalence: Who has the highest risk of being unemployed?

The model can used to predict the person specific exit and entry rates $\lambda_i(t)$ and $\mu_i(t)$. The first and the second answer are then given by the exit and entry probabilities in formula 6. The third answer corresponds to the lifetime unemployment of a person, which is the prevalence probability in formula 5. However, we can do even better if the last known state is u and the prediction is δt months to the future. We then use the transition probabilities in formula 4. Note that all these probabilities are implied by the same model.

We evaluate the model with a straightforward train and test set split. We investigate two different types of test sets, as plotted in the Figure 2:

1. Predict to the future: we train the model using 10000 persons in years 2013-2016 and take the year 2017 as the test set. We predict for the persons present in the training set, but require predictions for a future time.
2. Predict to new persons: we train the model using the 10000 persons in 2013-2016, and take another set of 10000 persons that the model has not seen and predict for them over the observation period 2013-2017.

The Receiver Operating Characteristic (ROC) curve illustrates the diagnostic ability of a binary classifier for different threshold values. We use the Area Under the ROC Curve (AUC) to evaluate the model performance. This is the probability that the model ranks a randomly chosen positive example higher than a randomly chosen negative example. For example, assume that the task is to predict who has the highest risk of escaping unemployment. We compute separately for every month in how many of the pairs where one person escaped unemployment and one did not, the exit rate prediction for the one who escaped was higher.

5.2 Predictive Accuracy

In Table 2 we present the time-stratified AUC of three different models, measured separately in the training set and the two tests sets. The Linear Model (LM) includes only the covariates but not the subject specific intercepts, the Linear Mixed Effects (LME) model is the statistical model, and Linear Machine Learning (LML) model is the machine learning version. We evaluate them on the three different prediction tasks: predict the risk of unemployment exit (Exit), unemployment entry (Entry) and the unemployment prevalence (Prevalence). The training set (Train) consists of years 2013-2016 with a sample of 10000 persons. The first test set (Test) consists of the year 2017 on the same persons, and the cold start test set (Cold) contains the years 2013-2017 for a different data set of 10000 persons.

Table 2. Time-Stratified AUCs of the models, evaluated in three different prediction tasks in one train and two test sets.

Model	Linear			Mixed Effects			Machine Learning		
	Train	Test	Cold	Train	Test	Cold	Train	Test	Cold
Exit	0.65	0.63	0.64	0.79	0.67	0.64	0.81	0.67	0.64
Entry	0.56	0.59	0.56	0.70	0.68	0.56	0.74	0.69	0.56
Prevalence	0.64	0.64	0.64	0.81	0.78	0.64	0.83	0.80	0.64

We make the following observations. First, the cold start prediction performance $(0.64, 0.56, 0.64)$ is the same for all of the three models. This is not suprising, since the LME and LML models are not able to learn the person specific intercepts for persons they have not seen. The rate prediction is based on the covariates only; the prediction formulas of these linear models are identical without the person specific intercepts and the learned parameters are very close to each other. The overall prediction performance is modest when we are forced to rely on the covarites only, but it is significantly better than random.

Second, the future test set predictions are improved significantly by using the LME and LML models $(0.63 \rightarrow 0.67, 0.59 \rightarrow 0.69, 0.64 \rightarrow 0.80)$ that include the person specific intercept. Part of this improvement could captured by including more detailed covariates in the model, but some part of it probably represents the person's characteristics that are difficult to measure. It is therefore useful to utilize the labour market histories with models that exploit this information. While predicting the exact timing of the transitions is still dificult $(0.67, 0.69)$, the overall unemployment prevalence can be predicted quite well (0.80) from the machine learning model.

Third, the training set provides overoptimistic performance measures in the LME and LML models that have a larger flexibility to fit the training set. This is probably due to the fact that the actual transitions to be predicted belong to training set. The differences on transition rates in the training set imply that we

actually know to some extent who had a transition and who did not, and this is reflected in the prediction accuracy.

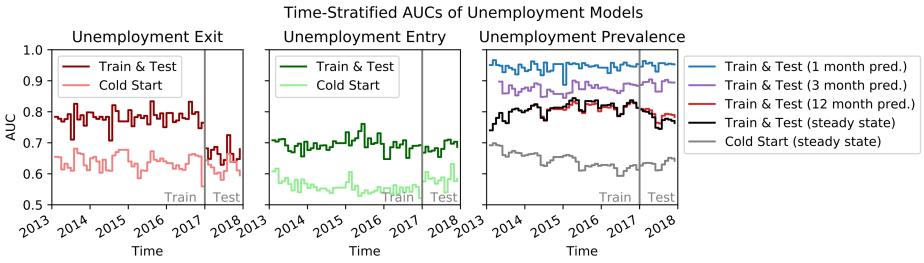

Fig. 3. Time-Stratified AUCs plotted separately for each month and prediction task.

Next we analyze the predictive performance over time, and whether the Markov Chain assumption enables even better predictions of who is unemployed at a given time. In Fig. 3, we report the LME model performance separately for every month. The predictive performance seems stable over time for unemployment exit and entry, taking into account that predicting unseen unemployment exists in the test set is considerably harder. The difficulty of predicting unemployment prevalence seems to correlate with the unemployment rate. When unemployment is high in general, the task is easier for the model that includes the person specific effects and more difficult for the model that doesn't.

In the unemployment prevalence plot, we included predictions that utilize the mathematical properties of the Markov Chain. We assume that we know the state 1, 3, or 12 months ago and use the transition probability matrix to calculate predictions for the future state. If one knows the past unemployment status, the task of predicting unemployment clearly becomes easier: it is very easy 1 month forward with AUCs in the range of 0.95, quite easy 3 months forward with AUCs up to 0.90, and rapidly more difficult with 12 months forward with AUCs of 0.80 being almost the same as the prevalence prediction performance.

5.3 Model interpretation

The linear models can be used to interpret the results. Every prediction is based on two sources of information: person's covariates (gender, work experience, age, level of education, field of education) and the person's labour market history. Some part of the person specific unemployment entry and exit rates are explained by the covariates and the rest can be inferred from the labour market history.

First we interpret the effect of covariates. We used the contrast sum coding when fitting the LME model, so that within a categorical covariate the parameters are constrained to sum to zero. Each parameter then estimates the risk of the covariate value, relative to the average of all values. We have plotted $\exp(\alpha)$, $\exp(\beta)$ and the implied prevalence $\exp(\alpha - \beta)$ on a logarithmic scale in Fig. 4.

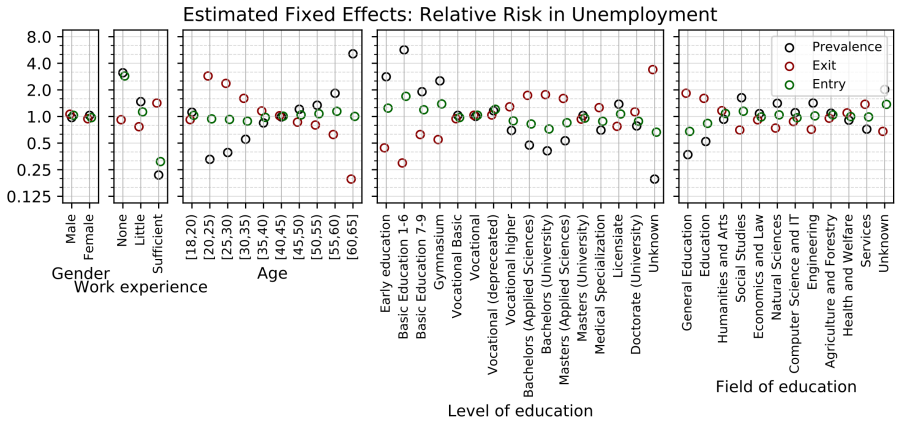

Fig. 4. Model parameters $\exp(\alpha)$,$\exp(\beta)$,$\exp(\beta - \alpha)$ corresponding to covariates $z_i(t)$.

For example, having age in the bucket [55,60) is predicted to result in two times the unemployent prevalence relative to the baseline.

It seems that gender is not a significant predictor. Having no work experience results in significantly higher unemployment entry and sufficient experience in significantly lower entry. The effect on unemployment exit is smaller, with experienced workers exiting faster. The result on unemployment prevalence is drastic: having no work experience predicts almost four times, whereas sufficient work experience predicts under a fourth, of the baseline. Younger adults have significantly better chances, and the persons close to retirement have significantly worse chances, of exiting unemployment. There is a slightly higher entry risk for both young and old people. As a result, the unemployment prevalence raises almost linearly with age and the effect for old or young people is again almost four times greater or smaller. Education is a reasonable predictor, where people with little education have higher prevalence because of both lower unemployment exit and higher entry, and people with higher education have significantly lower prevalence due to both higher exit and lower entry. The field of education is another reasonable predictor, with the person's degree implying a somewhat higher or lower exit rate from unemployment. These findings account for some of the differences in predicted transition rates.

Another component is the subject specific random effect, which we can predict from the model. We plot the pairs $\exp(u_i)$, $\exp(v_i)$ in Fig. 5. The estimated normally distributed random effects show some skewness and a significant negative correlation: a person with a higher unemployment exit rate tends to have a lower unemployment entry rate. The normal distribution mean vector is $\gamma = (-2.85, -1.79)$, and the estimated covariance matrix Σ implies a standard deviation 0.41 of the entry rate and 0.86 of the exit rate, with a correlation of -0.71. There is significant variation left between individuals even after accounting for their covariates.

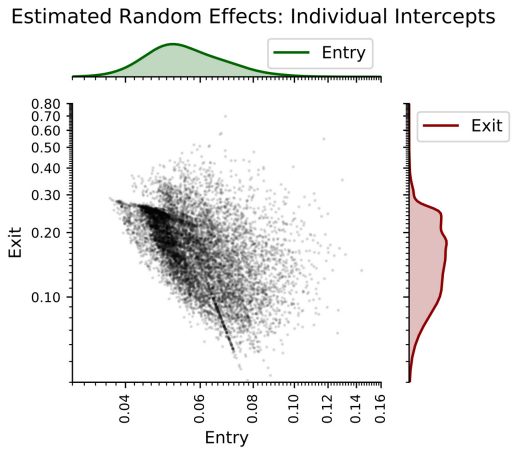

Fig. 5. Predicted random effects $\exp(u_i)$ and $\exp(v_i)$

.

6 Conclusion

In this study, we developed a model to predict the labour market state of a person. Our main focus was on prediction and we evaluated the model on three different prediction tasks: predict the risk of escaping unemployment, the risk of becoming unemployed, and the risk of being unemployed at any given time.

We used a Markov chain model with person specific transition rates. The transition rates were predicted by fitting three linear models to an unemployment registry: the simple linear model, the linear mixed effects model, and the linear machine learning model. We evaluated the models using time-stratified AUC on two test sets; one in the future and the second with unseen persons in the cold start setting. The person specific Markov chain assumption improves predictions significantly. The cold start problem is the hardest because one cannot use person history, and the models have a modest performance. On the other hand, predicting to the future for known persons is easier. It is still difficult to predict the exact timing of unemployment entry and exit, but we obtained good performance for the machine learning model of lifetime unemployment. Very good performance could be obtained to the near future given the last known state. The statistical model and the machine learning model result in similar predictions. The covariates have intuitive effects that are consistent with previous findings in the literature, but there is still considerable heterogeneity in the unemployment histories that can be used to improve predictions.

Our study has its own challenges and possibilities for future research. While registry data has many advantages, it is not necessarily reflective of the entire population in a given area. We investigated these biases in the appendix. Additional research could improve the predictive performance we have obtained with more detailed person information and non-linear models. Machine learning

has many models for high-dimensional data and non-linear relationships, but it would be useful to work with more detailed person level information to fully exploit their potential.

Appendix

Unemployment registry data has a number of potential biases if we want to generalize the results to the entire population. In that case, the training set should contain all 18 to 64 year olds currently residing in Varsinais-Suomi with their recurrent unemployment and employment spells. However, the unemployment registry is sampled monthly and contains only people who have been jobseekers at least once during the sampling.

The data set includes only people who have been unemployed

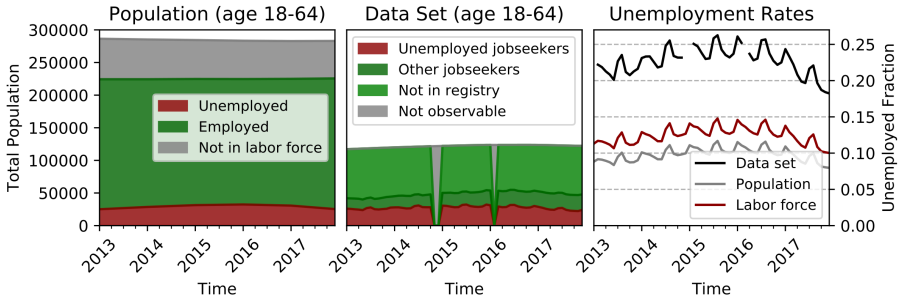

Fig. 6. Registry data set compared to the full population in Varsinais-Suomi.

Persons who have not been job seekers during 2013 to 2017 in the unemployment agencies are missing from the original data set because they do not have a registry entry. This is the case for people who have not been unemployed. We compared the data set to the official yearly statistics in Fig. 6, where we find that about 50% of the labour force in Varsinais-Suomi is missing. The unemployment exit rate is not biased, because every unemployed person is included. However, the baseline unemployment entry rate and prevalence are too high, because many people who are never unemployed are missing as negative examples. In other words, by definition we are analyzing unemployment among all people who experience unemployment at least once during the follow-up period.

It is still possible to estimate the true person specific rates from model predictions. Assume that the true unemployment entry rate is μ_i and the exit rate is λ_i for person i. Denote the length of a 'not unemployed' spell as T and the length of follow-up as t. The probability of missing from the data corresponds to probability of starting outside unemployment and remaining at that

state the entire time: $P(\{X_i(t) = 0\}_{t=0,1,\ldots}) = P(X_i(0) = 0)P(T > t)$. The data contains all of the 'unemployed' observations $P(X_i(t) = 1) = \frac{\mu_i}{\lambda_i+\mu_i}$ but the proportion of 'not unemployed' observations included is only $P(X_i(t) = 0) - P(\{X_i(t) = 1\}_{t=0,1,\ldots}) = P(X_i(0) = 0)P(T \leq t) = \frac{\lambda_i}{\lambda_i+\mu_i}(1 - e^{\mu_i t})$. Denote the observed odds of unemployment $\frac{\mu_i^*}{\lambda_i}$, which should be equal to the odds $P(X_i(t) = 1)/P(X_i(t) = 0)P(T \leq t) = \frac{\mu_i}{\lambda_i(1-e^{-\mu_i t})}$. This means we can solve:

$$\frac{\mu_i}{1-e^{-\mu_i t}} = \mu_i^* \tag{16}$$

We then obtain the true rate μ_i that produces the observed unemployment entry rate μ_i^*. With increasing follow-up $t \longrightarrow \infty$ we gather all samples.

The data set excludes some short spells

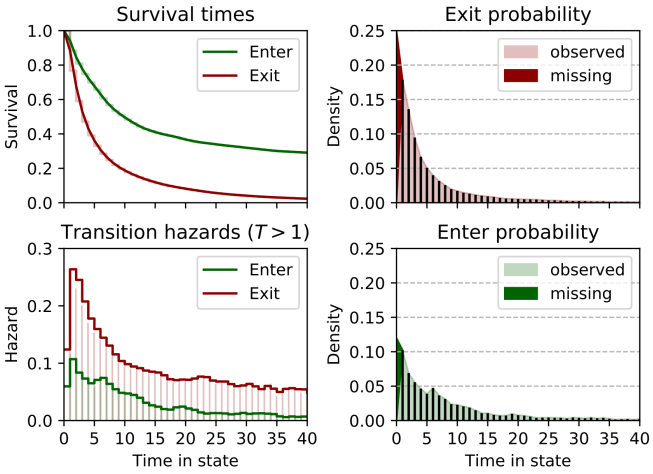

Fig. 7. Survival fuctions, hazards and probability densities of the unemployment (exit) and non-unemployment (enter) spell lengths, estimated from the first spell.

Spells shorter than one calendar month are undersampled because the registry status is recorded monthly. Such persons may enter and exit unemployment in between monthly measurements without being recorded. We estimate how many percent of spells are missing by calculating the Kaplan-Meier estimate $S(t) = P(T > t)$ of the first spell length T in Fig. 7. The second month hazard can be used to estimate the true first month hazard, as shown in the bottom left figure. For example, assuming that first month hazards should be 0.25 / month (exit) and 0.12 / month (entry), the percentage of spells that end in the first month should be $1 - e^{-0.25} \approx 22\%$ (exit) and $1 - e^{-0.12} \approx 11\%$ (entry) instead of the $1 - e^{-0.12} \approx 11\%$ (exit) and $1 - e^{(} - 0.06) \approx 6\%$ (entry) that were

observed. These spells are a small subset of the data set, and short spells do not meaningfully contribute to the total amount of unemployment

The data set includes some people who have a moved out

Finally, we have no knowledge of who remain or move out of the Varsinais-Suomi area. The data set may include persons who have moved out and are not at risk of being recorded in the unemployment registry. This bias can be estimated with a simple Monte Carlo simulation. From the government movement statistics in the years 2013-2017 (StatFin) we can calculate the migration rates within Finland. Each year on average 2.1% of the Varsinais-Suomi population moved into other economic areas, and 0.21% of the population in other areas moved into Varsinais-Suomi. We assume that the migration of people follows a Markov chain with the corresponding monthly transition probabilities. We then overlay the movement patterns generated from this Markov Chain into the data set as seen in the left of Fig. 8, and calculate the percentage of people that are outside Varsinais-Suomi each month in the right of Figure 8. This implies that about 6% of the samples at each time were probably outside Varsinais-Suomi.

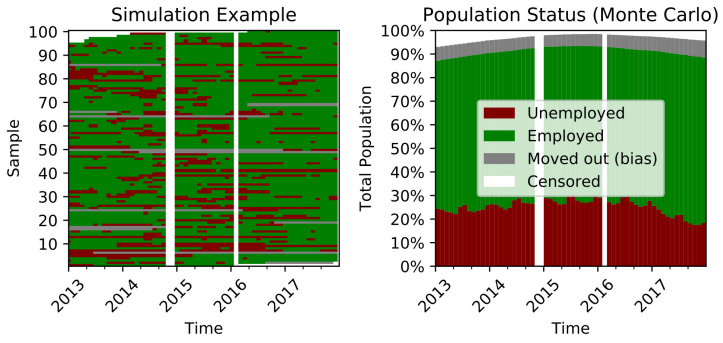

Fig. 8. The real world moving out bias is estimated with a monte carlo simulation.

References

1. Ernst, Ekkehard, and Uma Rani. "Understanding unemployment flows." Oxford Review of Economic Policy 27.2 (2011): 268-294.
2. Shimer, Robert. "Reassessing the ins and outs of unemployment." Review of Economic Dynamics 15.2 (2012): 127-148.
3. Ahn, Hie Joo, and James D. Hamilton. "Heterogeneity and unemployment dynamics." Journal of Business & Economic Statistics (2019): 1-26.
4. Honkanen, Pertti. "Odotelaskelmat työllisyyden, työttömyyden ja eläkeajan arvioinnissa." KELA Working Papers, No. 137 (2018).

5. Pedersen, Peder J., and Niels Chr Westergård-Nielsen. "Unemployment. A review of the evidence from panel data." Economics of Unemployment. Edward Elgar Publishing, 2000.
6. Wanberg, Connie R. "The individual experience of unemployment." Annual review of psychology 63 (2012): 369-396.
7. Kettunen, Juha. "Education and unemployment duration." Economics of education review 16.2 (1997): 163-170.
8. Ollikainen, Virve. "The determinants of unemployment duration by gender in Finland." VATT Discussion Papers, No. 316 (2003).
9. Kyyrä, Tomi. "Partial unemployment insurance benefits and the transition rate to regular work." European economic review 54.7 (2010): 911-930.
10. Rokkanen, Miikka, and Roope Uusitalo. "Changes in job stability: Evidence from lifetime job histories." IZA Discussion Papers, No. 4721 (2010).
11. Asplund, Rita. "Unemployment Among Finnish Manufacturing Workers. Who gets unemployed and from where?" ETLA Discussion Papers, No. 711 (2000).
12. Eriksson, Tor, and Jaakko Pehkonen. "Unemployment flows in Finland, 1969–95: a time series analysis." Labour 12.3 (1998): 571-593.
13. Peltola, Mikko. "Työmarkkinasiirtymät Suomessa. Työllisyyden päättymisen jälkeinen työmarkkinasiirtymien dynamiikka vuosina 1995-1999." VATT Discussion Papers, No. 360 (2005).
14. Heckman, James J., and George J. Borjas. "Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence." Economica 47.187 (1980): 247-283.
15. Flinn, Christopher J., and James J. Heckman. "New methods for analyzing individual event histories." Sociological methodology 13 (1982): 99-140.
16. Mühleisen, Martin, and Klaus F. Zimmermann. "A panel analysis of job changes and unemployment." European Economic Review 38.3-4 (1994): 793-801.
17. D'Amuri, Francesco, and Juri Marcucci. "The predictive power of Google searches in forecasting US unemployment." International Journal of Forecasting 33.4 (2017): 801-816.
18. Tuhkuri, Joonas. ETLAnow: A Model for Forecasting with Big Data–Forecasting Unemployment with Google Searches in Europe. No. 54. ETLA Report, 2016.
19. Katris, Christos. "Prediction of unemployment rates with time series and machine learning techniques." Computational Economics (2019): 1-34.
20. de Troya, Í. Martínez de Rituerto, et al. "Predicting, explaining and understanding risk of long-term unemployment." 32nd Conference on Neural Information Processing Systems. 2018.
21. Kütük, Yasin, and Güloğlu, Bülent. "Prediction of transition probabilities from unemployment to employment for Turkey via machine learning and econometrics: a comparative study." Journal of Research in Economics 3.1 (2019): 58-75.
22. Beyersmann, Jan, Arthur Allignol, and Martin Schumacher. Competing risks and multistate models with R. Springer Science & Business Media, 2011.
23. Tutz, Gerhard, and Matthias Schmid. Modeling discrete time-to-event data. Cham, Switzerland: Springer International Publishing, 2016.
24. Duchateau, Luc, and Paul Janssen. The frailty model. Springer Science & Business Media, 2007.
25. Cook, Richard J., and Jerald Lawless. The statistical analysis of recurrent events. Springer Science & Business Media, 2007.
26. Rausand, Marvin, and Arnljot Høyland. System reliability theory: models, statistical methods, and applications. Vol. 396. John Wiley & Sons, 2003.