Measuring non-Workers Labor Market Attachment with Machine Learning *

Nicolás Forteza and Sergio Puente

Banco de España

June 18, 2025

Abstract

Studying the Labor Market Attachment (LMA) for the non-working force is crucial for several economic outcomes, such as real wage or long-term non-employment. Official statistics rely on self-reported variables and rule-based procedures to assign the labor market status of an individual. However, this classification does not take into account other individual's characteristics, like variables related to reservation wages or the amount and type of job offers received, implying that estimates of the non-worker status could be biased. In this paper, we propose a novel methodology to measure the non-workers LMA. Using the Spanish Labor Force Survey (LFS), we define two groups (attached vs. non-attached), and estimate a probability distribution for each individual of belonging to such groups. To recover this probability distributions, we rely on unsupervised and supervised machine learning algorithms. We describe the differences between LFS unemployment, other measures of attachment in the literature, and our non-worker classification. We identify the instances in which our proposed methodology has a tighter relationship with measures like salaries, GDP and employment flows.

JEL Codes: J21, J82

*Corresponding authors: Nicolás Forteza (nicolas.forteza@bde.es) and Sergio Puente (sergio.puente@bde.es). We thank Juan Francisco Jimeno, Giorgio Topa, Manudeep Bhuller, an anonymous referee and participants at the internal seminar at Banco de España and ECONDAT 2024 spring meeting at King's College London for their useful comments.

1 Introduction

The measurement of labor market attachment is crucial for studying the economy. For instance, unemployment can affect labor market outcomes, such as wages (higher unemployment is typically related to lower wage growth, see for example Devereux and Hart (2006), Font et al. (2015) or Gertler et al. (2020)), whereas out of the labor force (OLF) people (for example, students, housekeepers, etc.) cannot. While the distinction between workers and non-workers is directly observable, to filter from the latter those who really are attached to the labor market is a more ambiguous matter. Traditionally, Labor Force Surveys (LFS from now on) rely on a set of questions to accomplish this distinction. With minor changes over the last decades[1], LFS in most countries define a non-worker as sufficiently attached to the labor market (i.e. as an unemployed) if the person is willing and available to work, and has actively looking for a job during the last weeks [2].

However, this binary and arbitrary distinction does not fully capture the true degree of labor market attachment of an individual (Jones and Riddell (1999)). Active search and willingness to work are not the only variables determining whether a person is really in the market or not. Other factors can play an equal or higher role here. For example, a person willing to work, and actively seeking, but with a reservation wage much higher than the equilibrium wage, could hardly affect labor market outcomes, despite being classified as unemployed. Same can be said regarding other factors, like the amount and type of job offers received, or job search intensity. Moreover, the distinction may not be binary, existing a gray zone between people attached and OLF, with intermediate degrees of attachment. The continuous aspect of the phenomenon is apparent when looking at transition probabilities into employment. For example, according to Eurostat, transitions from unemployment to employment during 2022 were 26% of the unemployed population in Spain. At the same time, transitions from inactive to employment were 11%, a lower figure, but not dramatically different. The probability of finding a job for those classified as inactive just because they are not immedi-

[1] Most of these changes operate at the definition of *active* search, see for example Garrido and Toharia (2004)

[2] People with an already signed, but not yet started, job contract, are included among unemployed, no matter if they searched or not

ately available to work raises to 33%, even higher than in the case of unemployed[3]. All these figures highlight the importance of considering intermediate attachment status, beyond the binary distinction unemployed-inactive.

In this paper, we propose a machine learning, data-driven methodology that aims to measure the labor market attachment (LMA) of those who don't work. Using Spanish LFS data, we train, in an adaptive manner, an estimator of LMA using a high dimensional dataset that contains multiple covariates related to job search, labor market outcomes, socioeconomic and demographic indicators, among others. Specifically, due to the novelty of the approach, we compare two possible novel ways of recovering the LMA probability distribution: an unsupervised vs. a supervised model. We first study what are the statistical properties of both LMA probability distributions. Then, we compare such distributions with other ad-hoc, binary methodologies across several dimensions contained in our dataset. Finally, we compare our LMA measure with other's LMA measure in various economic-related analysis: flows into employment, the correlation with salaries and with the economic cycle.

We find that our proposed LMA measure presents a continuous, multi-modal probability distribution. The implication is that each individual is attached to the labor market to some degree and that depending on the application, one could cluster individuals to form groups. In fact, when clustering such distributions into two groups, we observe many differences among our individuals attached to the market and the Official's and others' methodologies individuals that are attached (unemployed) in the labor market, in terms of demographic and *other* labor market related characteristics. Importantly, we observe that our supervised methodology outperforms other methods in predicting the probability of an individual to be employed the next year, used as a measure of attachment to the labor market. Lastly, we observe that the supervised methodology is more closely related to the economic cycle, even after excluding the Covid-19 quarters in the data.

There is a branch of the literature trying to fine-tune the distinction between non-working

[3]These patterns are also present in other countries: inactive people tend to have transitions to employment around half of those observed for unemployed, with some subgroups of inactive showing transitions equal or higher than those observed for unemployed.

people into those attached and those non-attached to the labor market. Most of the papers use ad-hoc, categorical classifications, adding new intermediate states. For example, Brandolini et al. (2006) add a third state to the unemployed and OLF states: people willing to work, and looking for a job, but whose last seeking activity occurred more than four weeks ago. They call this group "potential labor force". They find that this "potential" group is different both from unemployed and from inactive in terms of transitions to employment, and hence they propose to deviate from the usual binary definition. Overall, since Brandolini et al. (2006) work, all studies (Jones and Riddell, 2006; Lange and Kudlyak, 2014; Hall and Schulhofer-Wohl, 2018; Jones and Riddell, 2019) have used ad-hoc and rules-based estimators to create intermediate states and to describe the potential implications of the undermeasurement of unemployment.

Our contribution to the literature is multiple. We propose a data-driven approach to compute the degree of attachment to the labor market of an individual. Instead of a binary distinction (unemployed or OLF), we assign a degree of attachment of each individual to the labor market. This approach is agnostic to the researcher in terms of modeling, something that contrasts heavily with the related branch of the literature. Overall, the methodological part of the paper is well differentiated from other research papers, aiming to contribute to the literature related to the measurement of labor market states. We prove that using our preferred methodology of computing the LMA (the supervised one), one can better predict employment flows. This is of particular importance, because a higher LMA must ultimately reflect on higher probability of transition to employment, and viceversa. Lastly, we use those estimated LMA probability distributions, and by clustering the non-working people in two groups (attached vs. non-attached), we describe what have been the trends of attached and non-attached individuals in the Spanish labor market context.

The rest of the paper is divided as follows. In section 2 we comment on other papers that try to handle other type of biases when measuring the labor market states. Also, we explain briefly the ad-hoc distinctions proposed by other papers and what are the main implications. In section 3 we describe what are the dataset we are using to compute the LMA and its details. In section 4.1 we describe what is the main statistical framework for computing the

LMA measure and how it compares with other papers' estimators. We then develop specifically the general framework for both the unsupervised and the supervised methodology, providing an economic rationale and intuition behind it. Then, in section 5 we comment on the output of the modeling exercise: the LMA probability distribution, what are the differences of attached and non-attached workers vs. other methodologies and the potential economic consequences in flows into employment and prediction of salaries and the economic cycle. We stress the methodology with some changes to explore robustness in Section 6. Section 7 includes a first attempt to apply the methodology to other countries, which is limited due to data availability. We conclude with a discussion section in 8.

2 Related Literature

From a theoretical standpoint, what defines the labor force and hence, the unemployment and those attached to the labor market, depends on the assumptions on the researcher model. For instance, Lucas and Rapping (1969) argue that the labor force should be measured by those employed plus those who are unemployed but would accept work at what they regard as their normal wage rates[4]. In other words, unemployment is employment at job search since separations within the labor market are induced by the desire to search. In another work, Blanchard and Diamond (1992) assume that separations come from job destruction, and searching becomes an obsolete term, meaning that those "waiting" is what defines the unemployed side of the labor force. In other papers, Hall et al. (1970); Clark et al. (1979) stress this point: the distinction between unemployed and the out-of-the-labor-force (OLF) agents has always been ambiguous and sometimes arbitrary.

Not only the unemployment definition depends on the model's assumptions, but on the available data, often in the form of surveys. The ILO distinction of these groups using survey data can induce measurement error (since it depends on the survey design process and respondents circumstances) and temporal aggregation bias[5]. For instance, not including

[4]The normal wage rate is defined as *"wages in occupations in which he has formerly worked, wages of comparably skilled and aged workers, and so forth."*

[5]LFS flows provide snapshots of an individual's labor force status observed at a certain point in time. In practice, however, an individual may make multiple transitions between consecutive surveys. In consequence, LFS may miss some transitions and incorrectly include others. This phenomena, although tangent to this paper,

characteristics (for example, wage) of a potential acceptable job could be biasing the respondent answer to the job search question. Also, response error could imply the appearance of spurious labor transitions. For that matter, by means of using follow-up interviews of respondents, bias can be easily identified (Poterba and Summers, 1986). Other mechanism of error is the non randomness of missing survey respondents (Abowd and Zellner, 1985). Overall, applying these correction methods can bring to light the under-measurement of the unemployment rate (Feng and Hu, 2013; Ahn and Hamilton, 2022).

The distinction between the unemployed, the marginally attached and the OLF people, has important implications when studying the unemployment rate and its trends (Moffat and Yoo, 2015; Barnichon and Figura, 2016) or the unemployment and the business cycle (Elsby et al., 2015; Kroft et al., 2016; Shibata, 2022). From a methodological standpoint, Flinn and Heckman (1983) first defined and differentiated between unemployed and OLF individuals based on behavioral distinctions. They argued that in search theory, a key difference between unemployed individuals and those out of the labor force is that the former are actively engaged in job search activities, while the latter are not actively searching for employment and are effectively at a "corner" where they spend no time searching. This constituted the foundation for LFS across developed countries to measure unemployment. However, Jones and Riddell (1999) realized that, to properly measure unemployment, heterogeneity among nonemployed should be taken into account. They argued that discouraged workers for instance (those who want to work and are apparently unemployed, but are not actively looking) are seen as OLF, while in practice they are not. This raised questions about the behavioral distinctions between different groups of nonemployed individuals. But most importantly, this implied that some nonsearchers may still be attached to the labor force, challenging the traditional distinction between unemployment and nonparticipation based solely on job search activities. Since then, the inclusion or exclusion of marginally attached individuals among the unemployed has been studied. Brandolini et al. (2006) calculate that on average, in European countries about one-fifth of all people declaring seeking work, were left out of the labor force due to the LFS design: these people took their last step towards

has been studied in the recent decades (Darby et al., 1986; Shimer, 2012).

looking for a job more than four weeks before the interview. They called this group of people the "potential" labor force group. Using transition probabilities among european countries, they found that the nonemployed people would be better characterized by four distinct states: employed, unemployed, potentials and other OLF population. In a similar paper, Jones and Riddell (2006) define the marginally attached as the non-searchers who express a desire for work. Then, they analyze their transitions to employment within different multiple sub-categories of marginally attached (waiting for a job, nonwaiting, personal reasons for not being looking for a job, discouraged and other) and horizons (next month, two months ahead, etc.). They find that this heterogeneity among non employed is significant, meaning that these labor market states are distinct in terms of labor market behavior: the marginally attached are more likely to obtain employment than the nonattached, but less likely to obtain employment than the unemployed. Lange and Kudlyak (2014), using the Current Population Survey (CPS) data, construct six new different non employed labor force states based on the labor force status histories: unemployed recently employed, unemployed not recently employed, unemployed in the three consecutive months, OLF and recently employed, OLF and not recently employed, and finally OLF in the three consecutive months. They find that the new classification based on labor force status histories explains 25% more of the variation in job finding rates compared to traditional classifications. Hall and Schulhofer-Wohl (2018) construct up to 15 different non employed labor force states, exploiting detailed information of the CPS. They argue that accounting for this heterogeneity is crucial for measuring matching efficiency. Finally, Jones and Riddell (2019) taking into account those "wanting" to work as a fourth labor force status in US and Canada, decompose changes in the unemployment ratio into changes in the non employment rate, the labor force participation rate and the probability of unemployment given non employment. They find significant differences in the labor force attachment, the desire for work and search activity, indicating that decomposing the non employment group into a more precise classification yields better insights of the unemployment and labor force participation.

3 Data

We use the Spanish Encuesta de Población Activa (EPA) to calculate the LMA. The EPA is the correspondent Spanish LFS conducted by the Spanish National Institute of Statistics (Instituto Nacional de Estadística or INE). The survey provides quarterly national estimates with an approximate sample of 60.000 households and 160.000 individuals each wave[6]. Also, the survey provides a panel component, and individuals can be followed up to 6 quarters. Regarding the distinction between unemployed and OLF, EPA follows closely ILO recommendations. Hence unemployed individuals are defined as those who were engaged in active job search during the so-called "week of reference", and also are available to join a job immediately. Unemployment also includes those who already found a job, and are waiting to join in, no matter if they actively searched or not. On the other side, OLF are all other non-employed people. To construct our LMA measures, we use the EPA quarterly data from 2005 onwards.

The survey questionnaire design provides rich content that allows the researcher to observe multiple marginally attached labor force states of non employed people. For instance, the survey asks for questions regarding job search (whether the individual has been looking for any job and the reasons related to not be looking for), desire to work, time spent looking for a job, availability to start a new job and the reasons related to not being available, the previous job (if applicable) characteristics (sector, employee vs. entrepeneur etc.) and the subjective reasons of inactivity in the labor market. Panel A in the appendix shows the complete and detailed list of covariates and covariate interactions used for the statistical modeling purpose of this paper. By allowing for the LMA measurement models to observe all these variables, we are allowing for heterogeneity within the non employed to be a key ingredient of the LMA, in line with the works of Jones and Riddell (1999); Brandolini et al. (2006); Jones and Riddell (2006); Hall and Schulhofer-Wohl (2018). Also, including demographic and human capital related variables, such as age, sex, nationality and education, allows us to distinguish more granularly the labor force states. These kind of variables have been studied as important predictors of unemployment. For example, Ho and Tan (2008) find a non-linear relation

[6]The methodological documentation of the survey can be consulted here. It suffered some methodological changes in the recent decades, the last one happening in 2021, as is documented here. We control for this changes and harmonize all survey waves from the the first observed wave.

of education with unemployment, and Stephan Klasen and Silva (2021) study the effect of a set of covariates on the choice of participation among women. Moreover, these variables also have an effect on the duration of unemployment, and hence on the transition probabilities, see for example Fuchs and Weber (2017) or Krueger et al. (2014). Overall, the dataset comprises a total of 200 variables.

4 Methods

4.1 Framework

Suppose that an individual who is classified as unemployed, is also registered in the public employment office, perceives an unemployment benefit and spends less time looking for a job, compared with a similar peer, but without the benefit. Is the first person less attached than the comparison peer? Unemployment benefits can affect job search intensity, reservation wages, and the willingness to accept jobs of certain characteristics, yet the indivduals can still be classified as unemployed, see for example Feldstein and Poterba (1984) or Rebollo-Sanz (2012). On the other hand, there could be the case that an inactive (and thus, OLF person) has not been working for 4 months, but is willing to work if for any reason a friend offers him/her a job. In any case, the binary distinction of unemployed vs. OLF does not seem to fit well for these two individuals. In reality, one could say that they have different degrees of attachment to the labor market. We argue that there is some source of error given that this classification does not take into account the intensity of attachment, and its relation with other individual's characteristics: using the binary distinction may be an incomplete estimate of the true (possibly continuous) non-worker status.

Cited works of second paragraph in section 2 use researcher-generated ad-hoc rules, based on the available information of the correspondent LFS, to create additional non employed labor status. Specifically, the usual statistical model for characterizing different labor force states can be written as the following. Suppose we have a population of N individuals, each characterized by their labor force status Y_i and V survey responses X_{iv}. The labor force status Y_i is categorical, taking values from a set of categories $\{k_1, k_2, ..., K\}$, and follows a categorical

distribution. The estimator used comes in the form of a decision rule for determining the labor force status based on the survey responses. It can be expressed as follows:

$$\hat{f}_k(X_{iv}) = \begin{cases} 1 & \text{if } X_{i1} = v_1, X_{i2} = v_2, \dots X_{iv} = V \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $v_1, v_2 \dots, V$ are specific values for the survey responses corresponding to category k. In this context, the LFS uses three variables X_{i1}, X_{i2}, X_{i3}. These variables are: the person is not employed; the person is currently available for work; the person is actively seeking work, i.e. has taken specific steps in the four week period ending with the reference week to seek paid employment or self-employment or who found a job to start later. Then, the probability p_{ik} of individual i belonging to category k is determined by the decision rule

$$P(Y_i = k|X_{iv}) = p_{ik} = \hat{f}_k(X_{iv}) \tag{2}$$

and has only two outcomes (unemployed or not, OLF or not, etc.). Note that since the decision rule is deterministic and directly determined by the survey responses, there are no parameters to estimate. The estimated kth labor force state share is then

$$p_k = \frac{\sum_{i=1}^{n} p_{ik}}{n} \tag{3}$$

In this study, we argue that the above statistical model is useful but can be improved as follows. Suppose now each individual, instead of being part of some pre-defined labor force status, is attached to the labor market with some degree Y_i. If we believe that the LMA is a function of many observables (survey responses):

$$P(Y|X, \phi) = P_i = f(X_i, \phi) \in [0, 1] \tag{4}$$

X is the matrix of individual responses, and ϕ the parameter vector that optimizes f. Note that we are imposing the LMA to lie in between 0 and 1. The question here is how to approximate $\hat{f}$ in equation (4). Both in LFS and in the above-mentioned papers, $\hat{f}$ is estimated using some version of (1). We, instead, rely on machine learning tools to make an estimation actually based on a broad set of available individual information. Once we recover the probability density function of (4), we will eventually split the sample into K labor force status for comparisons purposes, in order to illustrate what are the main differences between our classification and LFS or other alternative methodologies.

4.2 Unsupervised

Suppose that an individual's survey responses hide some unobservable latent factor structure that resembles his true labor force status. For instance, variables belonging to heterogeneous groups (such as the spent time in searching a job, age and education or last time the individual worked and the availability to start working within the next two weeks) may be indicative of the same non-worker status. Within this context, dimensionality reduction seems a reasonable tool to extract the non-worker status from large individual data.The algorithm Latent Dirichlet Allocation (LDA), proposed by Blei et al. (2003), known for being used with text data for recovering topics within a set of documents (news, speeches, etc...), seems a good fit for this purpose. We use it to recover latent employment status of individuals. The algorithm works as follows. The survey responses (covariates) X is a IxM matrix, with I individuals and M covariates. Assuming again K possible labor status:

- For each labor force status $k \in K$, draw $\beta_k \sim Dir(\eta)$. This provides the probability of a survey response or covariate appearing in labor force status k, and η is its prior.
- For each individual $i \in I$, draw $\theta_i \sim Dir(\alpha)$, where α is the prior of an individual labor force status distribution.
- For each survey response m of individual i in X_i:
    1. Draw the labor force status assignment $z_{im} \sim Mult(\theta_i)$
    2. Draw the observed response $X_{im} \sim Mult(\beta_{z_{im}})$

The posterior distribution is:

$$P(z, \theta, \beta | X, \alpha, \eta) = \frac{P(z, \theta, \beta | \alpha, \eta)}{P(X | \alpha, \eta)} \tag{5}$$

Suppose we fix $K = 2$. Since we are only interested in the probability of an individual to being part of one group (attached to the labor market) or another (non attached), if we take all individuals distribution over both labor force status matrix θ, the model returns the probability of being attached to the labor market:

$$P(\theta | X, \alpha, \eta) = P(Y | X, \phi) \tag{6}$$

Equation (5) is estimated with variational Bayesian methods since it is intractable. This model assumes that we do not have any prior beliefs in what is the attachment of each individual. We also don't know the extent to which each variable within X_v impacts the decision of being attached or not. For that reason, we fix both hyperparameters α and β to 0.5. Also, the decision of the temporal coverage of the sample has some implications. As our sample contains multiple years and quarters, we use a moving window of one year of historical data (from $t - 3$ to t) to estimate equation (6). And then, we use it to retrieve the status of all individuals, based on their individual information at t. Figure 1 shows the setting we use to fit our model for individuals in time t.

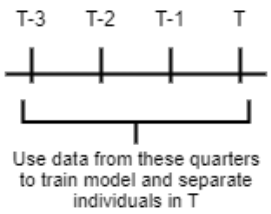


Figure 1: Unsupervised model training scheme.

This approach entails two advantages. First, by using exactly four quarters in the training setup, we get rid of labor market seasonality. It is well documented that seasonality and

calendar effects have a large impact in the Spanish labor market (see for example (Conde-Ruiz et al., 2019)). Second, by using only one year of data, we allow the model to react in a relatively fast way to changes in data patterns over time.

4.3 Supervised

Instead of recovering some hidden latent factor structure, we can approximate the LMA function by minimizing some loss function. More specifically, the main idea is to recover an individual's probability of being employed in the future, so that the predicted probability resembles the LMA. Following the previous subsection notation, suppose we assume the function f is a linear combination of X such that:

$$P(Y|X,\phi) = p(X,\phi) - \lambda\|\phi\|_1 \tag{7}$$

To estimate ϕ^*, we can minimize the following log-loss function:

$$l(\phi^*) = \max_{\phi}\left\{\sum_{i=1}^{I}[y_i log(p_i) - (1-y_i)log(1-p_i)] - \lambda\sum_{j=1}^{J}|\phi_j|\right\} \tag{8}$$

where y_i is a dummy indicating if individual i is employed in any quarter between $t+1$ and $t+4$, and p_i is the logistic function $\frac{1}{1+\exp(-\phi X)}$, which measures the probability of being employed the next year. Finally, λ is the regularization parameter that controls the strength of the L1 penalty. A larger λ value leads to stronger regularization and more feature weights being set to zero. Given the high-dimensional setting we are into, a regularization parameter makes sense since we want to avoid overfitting and get rid of noisy covariates that would also hinder the interpretation and analysis. The decision of predicting individual employment the next year ahead is, again, because of seasonality issues: using only one quarter ahead would imply most of the action being captured by quarter dummies. We instead prefer to give more weight to individual information. Overall, the researcher can utilize some other way of estimating ϕ^* (the first part of the right-hand side of Equation 7), such a more sophisticated model (neural nets, for instance).

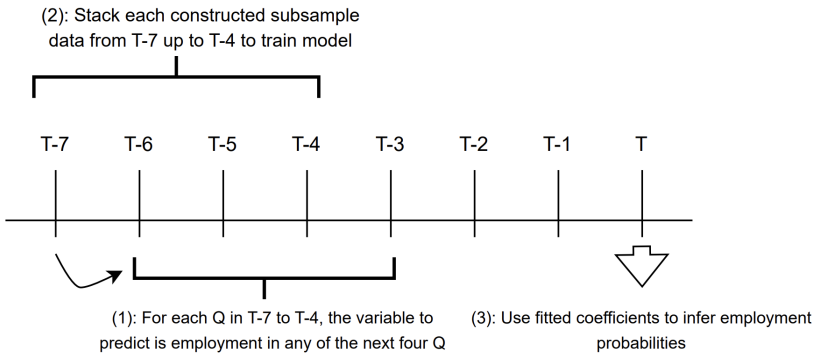


Figure 2: Supervised model training scheme.

In figure 2 we can see the training setup for the supervised model. We still use four quarters, and the target variable is the probability of being employed in any of the next four quarters ahead, so at a given point in time t, we use data from $t-7$ (where target variable is employed or not from $t-6$ to $t-3$) to $t-4$ (target variable is employed or not from $t-3$ to t) to fit the model. We apply the estimated coefficients to individuals at t to then get the probability of being employed at any quarter of the next year. This procedure can be better understood step by step. For each t:

- We split the data from $t-7$ up to $t-4$ into 70% training set and 30% test set. Doing a hyperparameter search over a grid of λ, we observed that on average, 0.5 as a L1 penalty minimizes equation 1.
- We estimate the model (coefficients) on the $t-7$ to $t-4$ data subsample.
- If we have available coefficients (note that we can't infer for the first two years), we calculate probability of being employed in the next year ($t+1$ to $t+4$) for observations in t.

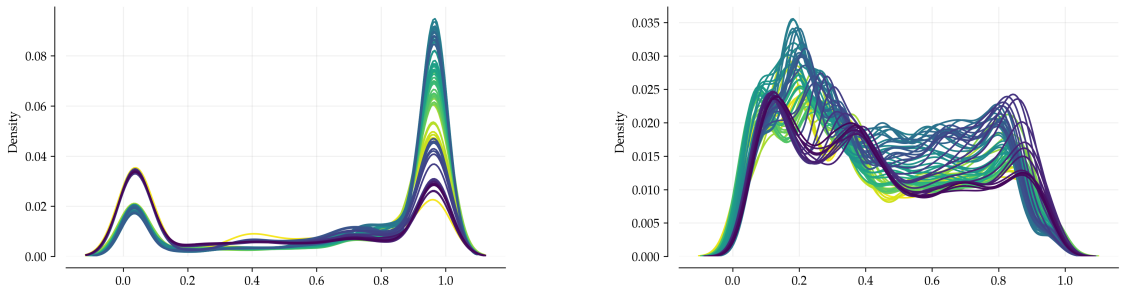
Note that with this setup, we never use future information, so we can calculate individual probabilities in real time.

5 Results

5.1 Labor Market Attachment

In Figure 3 we observe the fitted LMA probability distributions of each model: unsupervised (left) and supervised (right). The darker color represents the first year of observation (2006Q1 for the left figure and 2007Q1 for the right figure), and the lighter color represents the last quarterly inferred distribution (2023Q1).

Figure 3: Labor Market Attachment Probability Distribution



(a) Unsupervised (b) Supervised

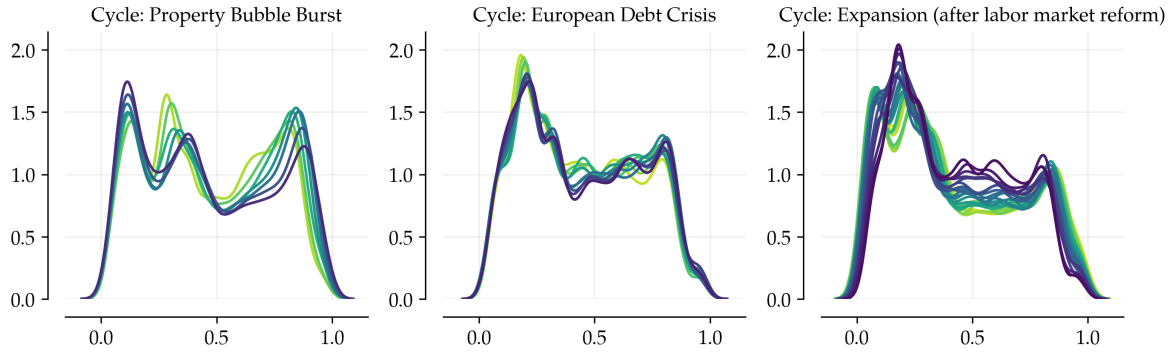
Notes: Each panel shows the yearly probability distribution of LMA. The x-axis represents the level of labor market attachment, ranging from 0 to 1, while the y-axis shows the density of the distribution. The lighter color represents the first year of observation (2006Q1 for the left figure and 2007Q1 for the right figure), and the darker color represents the last quarterly inferred distribution (2023Q1).

In the unsupervised model expressed in Equation (5), we set the hyperparameter $K = 2$, so a single probability defines completely the distribution over groups. Looking at Figure 3a, we observe this probability follows a bimodal distribution, with some variation over time due to the inclusion of up to four quarters in each model training and inference as seen in Diagram 1. The lower-magnitude mode proximal to 0, is indicative of minimal labor market attachment (or inactivity). The distribution near 0 exhibits a slight diminution in magnitude coupled with increased dispersion, potentially indicating a more nuanced spectrum of low attachment states. The higher-magnitude mode approaching 1, indicates robust labor market engagement (or being unemployed, looking for a job, waiting in the transition in between one job and another, etc.). The peak near 1 demonstrates increased density and reduced variance, suggesting a consolidation of the highly-attached labor force segment. The region between the modes maintains a relatively consistent, albeit low, density. This could represent

a quasi-stable transitional state with intermediate labor market attachment.

Moving to Figure 3b, we observe multimodal LMA probability distributions. The density curves demonstrate multiple peaks, suggesting a highly heterogeneous underlying population in terms of the probability of being employed in the next year, as pictured in Diagram 2. This is reasonable since we are not imposing any restriction in Equation 8. Actually, note that areas closer to 1 represent the probability of being employed the next year, as pictured in Diagram 2. A prominent mode is evident in the 0.1-0.3 range on the x-axis, exhibiting the highest density values. Secondary modes are discernible around the 0.5-0.6 and 0.7-0.8 ranges, though with lower densities. The evolution of the curves suggests significant shifts in the underlying distribution over time: the primary mode (0.1-0.3) appears to gain density and shift slightly rightward in later periods. Interestingly, both tails of the distribution (near 0 and 1) show consistent low density across all time periods. The frequent intersections of density curves from different time periods suggest a dynamic system with complex transitional properties, potentially indicating frequent reallocation or reclassification within the measured population: a highly segmented labor market, with distinct subpopulations exhibiting different levels of attachment. Finally, the fitting process of the model following the structure of Diagram 2, can be observed in Figure 15 in the Appendix. It shows the ROC AUC curve metric for different subsamples during the training phase and the inference phase for the out-of-sample individuals (the assignment of the LMA score to each individual).

Figure 4: Supervised LMA during Economic Cycles



Notes: Each panel shows the yearly probability distribution of the computed supervised LMA. The x-axis represents the level of labor market attachment, ranging from 0 to 1, while the y-axis shows the density of the distribution for different economic periods of the cycle.

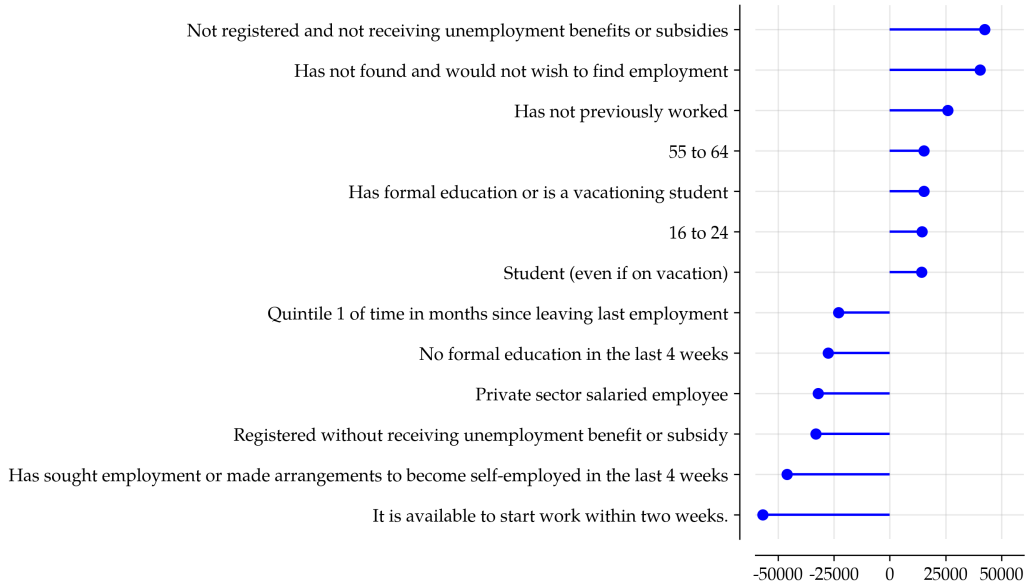
Given the changing nature of the distribution estimated with the supervised method, it is interesting to see if some pattern over the business cycle appears. In this connection, Figure 4 splits the distributions into three different sub-periods: The first part of the financial crisis (2008Q3 to 2010Q1), its second dip related to the ensuing debt crisis (2011Q2 to 2013Q3) and the posterior recovery (2013Q4 to 2019Q4). In all three, there is a prominent mode at low degrees of attachment, probably capturing people out of the labor force, or some long-term unemployed with low employability. In contrast, a mode in higher levels of attachment is more apparent in the first graph, capturing people losing the job during the recession, and hence still keeping a strong attachment to the labor market. The middle graph shows that this group is less important, probably suggesting a drop in employability over long non-employment spells. Finally, the last graph also shows a reduced high-attachment group, this time reflecting that the highest attached people are finding jobs, leaving the less attached ones overrepresented among nonworkers.

Putting together the two approaches, the results suggest that non-employed population can be classified into two (relatively disjoint) groups according to their individual characteristics. But once we switch the focus to the probabilty of reentering employment, the clear partition disappears, with many individuals lying on intermediate probabilities. This supports the idea of looking for intermediate states, followed by the literature on marginal attachment cited before.

What are the drivers of such distributions? In Figure 5 we observe the most important variables determining the Unsupervised method output of Figure 3a. In LDA, group-variable distributions represent the probability of each variable appearing in a given group (recall $K = 2$, so there are two groups). These probabilities indicate how strongly a variable characterizes a group.

The plot displays a series of variables along the y-axis, with their corresponding importance scores represented by horizontal blue lines and dots on the x-axis. Positive values indicate the most important variables related to the non-attached group (or non-workers with a low degree of attachment) and viceversa. The first result we find is that the variables used in the

Figure 5: Relative Importance of Variables in Unsupervised Method



Notes: The plot displays a series of variables along the y-axis, with their corresponding importance scores represented by horizontal blue lines and dots on the x-axis. Positive values are variables related to lower attachement and negative values are variables indicating higher attachement to the labor market.

standard ILO classification are also very important here. For example, "Has not found employment and would not wish to find employment" is strongly related to the non-attached group, whereas active search and availability to work are the two most important variables determining attachment. This is a striking result, because we are not forcing the model by any means to replicate ILO classification. Yet, these variables are not the whole story. Some other variables appear, possibly related to the reservation wage: Students tend to be classified as non-attached (they probably need a higher wage to compensate for abandoning studies), while people not receiving an unemployment benefit tend to be attached. There is another set of variables which can be related to the amount and quality of job offers received: People too young or old, as well as new jobseekers, tend to be non-attached. On the other hand, recent work as a private employee positively affects attachment.

For the Supervised case, Figure 6 plots the most relevant estimated coefficients of Equation (8) that explain either a higher or lower labor market attachment, together with the three main variables that define unemployment. Regarding the latter, they mostly have the expected sign, with some exceptions in some years (like active search during the financial

Figure 6: Top Coefficients of Supervised Method

Notes: This graph presents a time series analysis of various factors influencing labor market attachment from 2007Q1 to 2023Q1. The coefficients represent the strength and direction of each factor's impact on labor market attachment, with positive values indicating higher attachment and negative values indicating lower attachment.

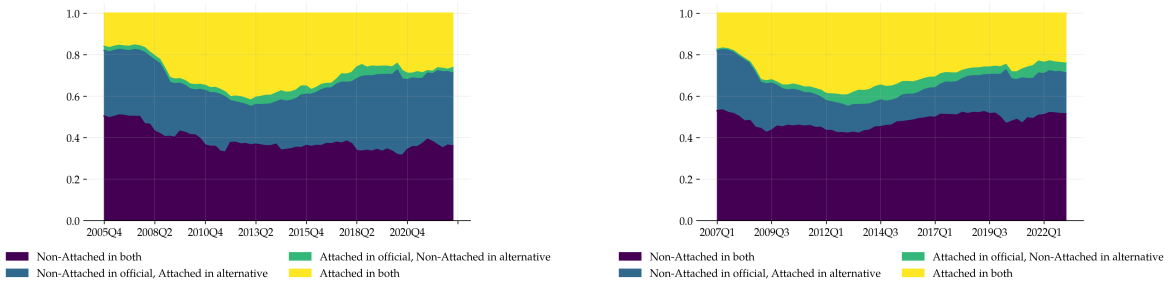
crisis). However, their magnitude is dwarfed by other coefficients. Having lost the last job less than one month ago is an important determinant of attachment, which indicates that recent job separations tend to be followed by relatively quick re-employment. In addition, people who complete their tertiary education have much greater attachment, compared to other first-job seekers. On the other side, older people have a lower attachment, being this effect quite volatile, and somehow less intense in recent years. Again, these results highlight the importance of other variables measuring reservation wages or employability, beyond the questions used for unemployment definition. The time series also reveals some macroeconomic patterns. There appears to be a general compression of coefficients during the 2008-2009 financial crisis, followed by increased dispersion afterward. Additionally, the period around 2018-2020 shows increased volatility across several groups, possibly reflecting labor market disruptions preceding and during the early COVID-19 pandemic.

5.2 Labor Market Attachment vs. Official Statistics

How does our labor market attachment measure differ from official statistics? Note that we do not make a binary classification of non-working population into attached and OLF subpopulations. Instead, we say that a non-worker is attached to the labor market to some degree. However, to make comparisons with official statistics, we should use official classifications as a benchmark. For instance, looking at Subfigure 3a, the bimodality of all quarters' distributions allows us to separate almost perfectly the non-working population into unem-

ployed (those with LMA close to 1) and OLF (close to 0). Thus, 0.5 is a threshold that can separate both groups. However, for the supervised case, since it presents a multimodal form, with substantial mass in middle values, the threshold selection is not so obvious. Graph 17 in the appendix shows several moments of the distribution. We set the threshold to be 0.4, which is the consistent median across time of the inferred supervised LMA.

Figure 7: Comparison with Official Statistics



(a) Unsupervised

(b) Supervised

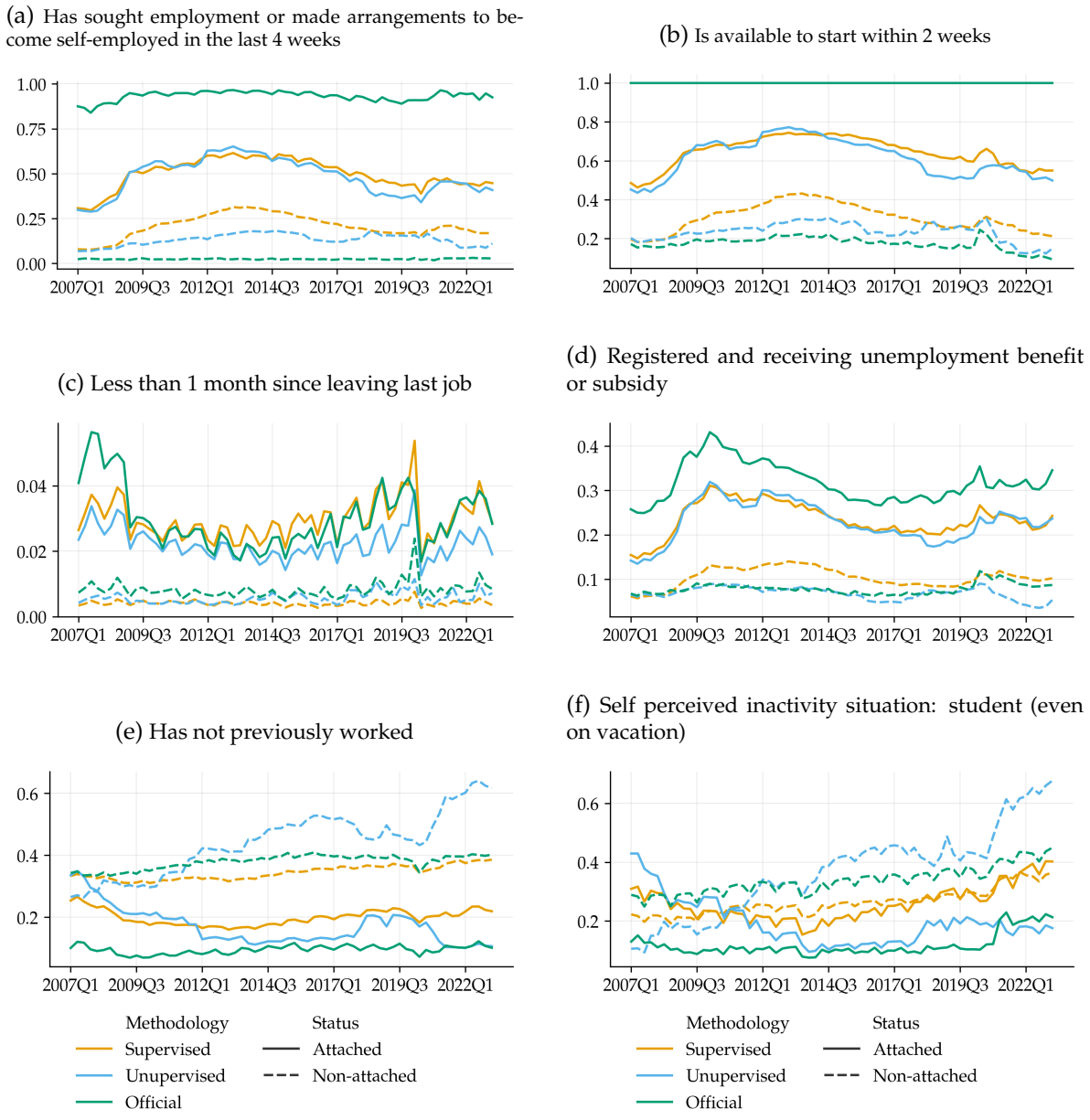
Notes: Each panel shows the relative size of each classified subpopulation. The x-axis shows time, while the y-axis shows the size of all four subgroups. The yellow color shows people both classified as attached (unemployed in the official classification and attached by our respective methodology), the light green color shows people unemployed in the official statistics but classified as OLF by our respective methodologies, the dark turquoise color shows OLF people by the official statistics but attached by our respective methodologies, and purple color shows OLF people classified by official statistics and non-attached by our respective methodologies.

One could interpret Figure 7 as a confusion matrix. Yellow and purple color shows whether our proposed method coincides with official statistics. If our methodologies delivered the same result as official statistics (official unemployed people being detected as attached to the labor market and official inactive people being classified as non-attached to the labor market by our methodologies), we would observe only yellow and purple areas. However, we observe a common pattern across both proposed methodologies: there's a significant portion of officially classified people as OLF that actually are attached to the labor market. We add people to the unemployed group, coming from the officially classified OLF subpopulation (dark turquoise group). In other words, as stated in section 4.1, there could be some omitted variable bias that could be distorting the actual level of unemployment in the economy. We will dig deeper into this question in the next section. Also, the light green color shows people we add to OLF, coming from official unemployment.

Another way of comparing our methodologies with the official one is by looking at individ-

ual characteristics. Specifically, in Figure 8 several important variables used to infer the LMA across all different methodologies (official statistics, the unsupervised method and the supervised method) and subpopulations (attached (unemployed in official) and non-attached (OLF in official)). The y-axis shows the % of subpopulation represented in each methodology-classification-variable bracket.

Figure 8: Characteristics of Attached vs Non-Attached non-workers



The first two graphs show two characteristics directly related to the definition of unemployment. Graph (a) plots the proportion of people who have sought employment or made

arrangements to become self-employed in the last four weeks. Here, the official unemployed consistently reports the highest percentage, hovering around 100%[7], while supervised and unsupervised methods suggest that active search is relevant for attachment, but not as important as in the case of official unemployment. The same happens in graph (b), which plots availability to work. Graph (c) shows that having lost the last job very recently is very important for attachment, especially in the supervised case, with the largest distance between attached and non-attached people. In graph (d) we see that being registered as a job seeker in the public employment services, and earning an unemployment benefit, is strongly related to unemployment. However, in our methodologies -especially in the supervised one- the distance between attached and non-attached is smaller. This suggests that benefits earned by some job seekers put their reservation wage so high that they are effectively out of the market. Graph (e) shows that the absence of previous experience is very important for the unsupervised case, especially in recent years, but not so for the others. The same is true for self-perceived students (graph f), with the additional interesting result that this variable presents no difference between attached and non-attached. This latter results possibly reflects that students can receive relevant job offers, which put them as attached as others, no matter the value of other variables.

5.3 LMA for different Socio-economic Groups

In light of Figure 8, another way to present the results is to compare the attachment between different methodologies with a focus on specific socio-economic groups. In this subsection, we will focus on three of them: Students, housekeepers, and gender differences.

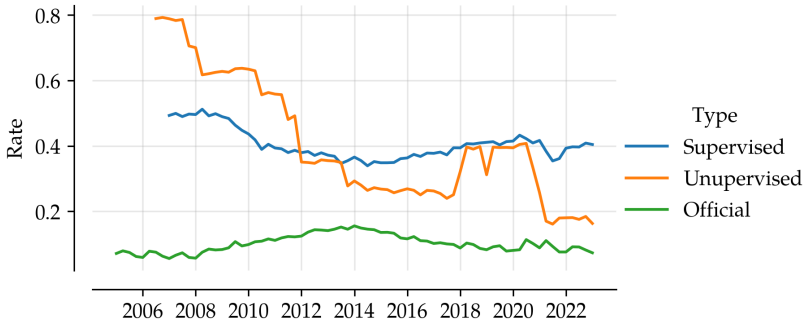
Starting with students, Figure 9 plots the average attachment computed by the three competing methodologies, but restricting the sample to only students[8]. Among them, few are classified as unemployed, with some minor increases during the financial crisis. This means that students tend to avoid an active job search throughout the business cycle. However, the picture changes when we look at our two attachment estimations. Both show a substantially

[7] It is not 100% because people who already found a job, and will join in shortly, are classified as unemployed, no matter if they looked for a job or not

[8] Defined as those who answered...

larger degree of attachment, compared to official unemployment. In the unsupervised case, the attachment tends to decline over time, suggesting that students are less similar to others in recent times. The supervised methodology delivers a more stable attachment over time, but with a marked cyclical pattern. This possibly reflects the ability of students to find jobs despite not searching for them, which is more relevant when economic growth is intense.

Figure 9: LMA: Students



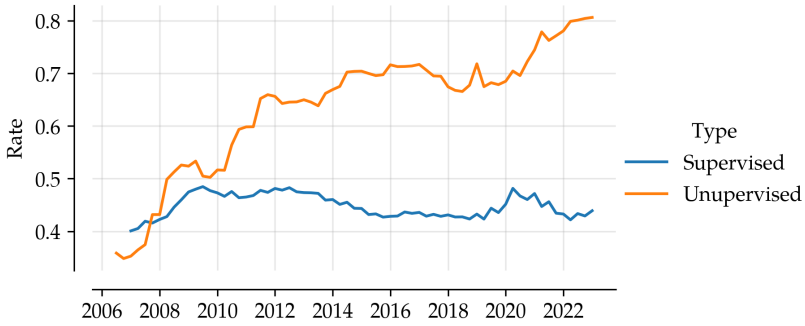
Notes: LMA is computed for each subpopulation, non-working group. For the Official classification, attachment is equal to unemployement (in the form of the binary distiction), and for the supervised and unsupervised case, attachment is equal to our fitted LMA probabilities.

Another interesting group is those reported as inactive due to being involved in housekeeping activities. For this group, the attachment measured by unemployment is zero by definition. Despite that, our two measures show medium to high attachment, as shown in Figure 10. In the unsupervised case, attachment grows over time, with a reversed interpretation compared to students: Housekeepers tend to be more similar to other non-employed people in recent times. The supervised method also presents reversed results with respect to students: Housekeepers' attachment is higher during downturns. The most sensible interpretation here comes from the need of some households to get a second wage earner when the main earner has difficulties in keeping the job or the wage. After the COVID crisis, attachment in this group is particularly high, probably reflecting people temporarily becoming housekeepers for a short period of time due to health issues.

Regarding gender differences, Figure 11 plots the unemployment rate by gender, together with the equivalent calculation for our two methods [9]. As can be seen, our two measures

[9]Specifically, for each non-working people in the sample, we compute the two attachment measures, and give them the corresponding probability of being attached. Then, we compute the weighted sum of all these proba-

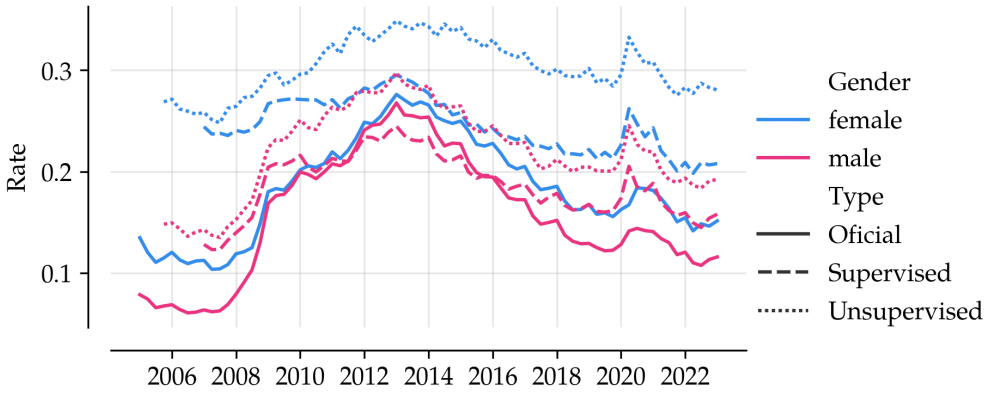
Figure 10: LMA: Housekeepers



Notes: LMA is computed for each subpopulation, non-working group. For the supervised and unsupervised case, attachment is equal to our fitted LMA probabilities. For the Official definition, housekeepers are out of the labor force; so that is why we cannot compare this subpopulation with the Official one.

of attachment are higher than the official one, coherently with Figure 7. However, when looking at the two genders separately, the main differences between methodologies come from women, especially at the start of the sample. This result points toward a LMA for women higher than job search intensity suggests. The corollary is that the steady increase in labor market participation for women during recent decades would have been less intense, as some of these women were actually attached despite being classified as inactive.

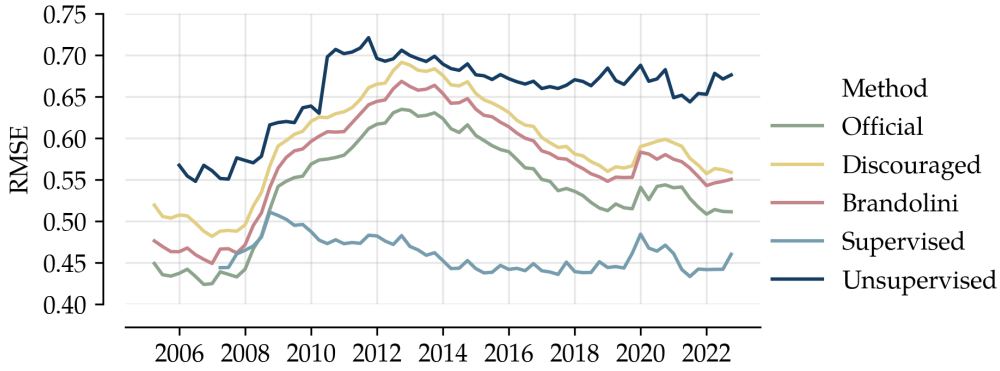
Figure 11: LMA: by Gender



Notes: LMA is computed for each subpopulation, non-working group. For the Official classification, attachment is equal to unemployement (in the form of the binary distiction), and for the supervised and unsupervised case, attachment is equal to our fitted LMA probabilities.

bilities to get a measure of total number of attached people. Finally, attachment ratios similar to unemployment rate are computed by dividing attached over the sum of attached and working people

Figure 12: Predicting Employment in $t+1$: RMSE



Notes: Sample restricted to non-retired and non-employed. The *discouraged* classification adds w.r.t the *official* classification the following: OLF people not looking for a job, available to work and willing to work. The *Brandolini* classification restricts w.r.t the *discouraged* classification to being looking for a job in the previous period. The standard errors are three orders of magnitude smaller than the values shown in this Figure; therefore, they are visually indistinguishable, suggesting high precision in the estimates across all methodologies.

5.4 LMA and its Economic Consequences

As we argue that our methodologies takes into account hidden factors that are ommitted in other measured employment states' methodologies, we'd expect that our LMA measures outperform other methodologies in predicting employment flows. For that purpose, we compute the root mean squared error (RMSE) of being attached to the labor market and the actual employment state in $t+1$.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

In (9), y_i a binary indicator of non-worker i observed in t but being employed in $t+1$, and $\hat{y}_i$ the classification outcome of any of the compared methodologies; e.g. $\hat{y}_i$ is indicator of non-worker i observed in t classified as attached to the labor market in t. Note that the official and other methodologies use binary data: they say whether an individual is unemployed or not, so $\hat{y}_i$ is a binary indicator. Our methodologies recover a continuous indicator of LMA, so that $\hat{y}_i$ will be between 0 and 1. For each quarter, we sum all individual RMSEs, weighting them by the population weights. Figure 12 plots such outcome.

Figure 12 compares five distinct methodologies: Official, Discouraged, Brandolini, Super-

vised, and Unsupervised. Each methodology represents a different approach to classifying individuals' labor market status. The *discouraged* classification takes the *official* unemployed, and adds as attached the following set: OLF people not looking for a job, available to work and willing to work, as described in Cuadrado and Tagliati (2018). The *Brandolini* classification restricts the *discouraged* classification to being looking for a job in the previous period (Brandolini et al., 2006). A notable trend across all methodologies is the general increase in RMSE from 2006 to around 2013-2014, followed by a gradual decline towards 2022. The Official methodology consistently shows lower RMSE values compared to the Discouraged and Brandolini approaches. This suggests that the official classification may be more predictive of future employment status. The Discouraged methodology, which expands on the official classification by including out-of-labor-force (OLF) individuals who are available and willing to work but not actively seeking employment, consistently demonstrates the highest RMSE among the non-machine learning approaches. This indicates that including discouraged workers may introduce more variability in predicting future employment status. The Brandolini methodology, which further restricts the Discouraged classification to those who were looking for work in the previous period, shows RMSE values intermediate between the Official and Discouraged approaches. This suggests that considering past job-seeking behavior may improve predictive accuracy compared to the broader Discouraged category. Interestingly, the Supervised methodology exhibits, by far, the lowest RMSE values. This implies that machine learning techniques may offer superior predictive power in forecasting employment transitions[10]. Conversely, the Unsupervised methodology consistently produces the highest RMSE values.

Another way of looking into potential consequences of using our LMA measure is to assess its relationship with other macroeconomic variables (Moffat and Yoo, 2015; Barnichon and Figura, 2016; Elsby et al., 2015; Kroft et al., 2016; Shibata, 2022). Table 1 presents a comprehensive analysis of the relationship between various LMA methodologies and key economic indicators, specifically total salaries, market salaries, and real GDP growth rate. The analysis

[10]The supervised model targets transitions to employment, so it not surprising that it performs the best here. However, note that this result is not driven by any kind of overfitting problem, as the model is estimated in real time, i.e. using only past information to forecast future transitions.

employs five distinct methodologies for classifying labor market attachment: Official, Discouraged, Brandolini, Supervised, and Unsupervised, as in Figure 12. In all the five, we calculate equivalents of unemployment rate, by taking the ratio $LMA/(LMA + Employment)$[11]. For each economic indicator, the table displays two sets of regression results: one including the full dataset and another excluding the COVID-19 period in the bottom row of each regression, allowing for comparison of the methodologies' performance under normal and extraordinary economic conditions. Across all methodologies and economic indicators, the coefficients are predominantly negative and statistically significant, indicating an inverse relationship between the measured labor market attachment and economic outcomes, in line with other studies (Cuadrado and Tagliati, 2018; Font et al., 2015). The R-squared values provide insight into the explanatory power of each model. For total salaries and market salaries, the Official, Discouraged, and Brandolini methodologies exhibit similar R-squared values (around 0.73-0.78), suggesting comparable performance in explaining variation in wages. The Supervised methodology, however, shows notably lower R-squared values (0.52-0.59), indicating less explanatory power. Interestingly, for the real GDP growth rate, the Supervised methodology demonstrates the highest R-squared values (0.87 with full data, 0.74 excluding COVID), substantially outperforming other methodologies. This is not surprising, as transitions to employment are what is relevant for GDP growth, whereas other considerations, maybe better captured by other methodologies, can have an added role in wage formation. The Unsupervised methodology consistently shows strong performance across all economic indicators, with R-squared values comparable to or slightly lower than the traditional methodologies for salary outcomes, and second only to the Supervised approach for GDP growth. When comparing the full dataset to the results excluding the COVID-19 period, we observe a general decrease in R-squared values across all models. This reduction is particularly pronounced for the GDP growth models, suggesting that the relationship between labor market attachment and economic growth may have been disrupted or altered during the pandemic. The coefficients for the Unsupervised methodology are consistently larger in magnitude compared to other approaches, indicating that this method may be capturing a broader or more sensitive measure of labor market attachment. In conclusion,

[11] For our methods, we calculate LMA as the weighted sum of our computed probabilities of attachment

while traditional methodologies (Official, Discouraged, and Brandolini) perform similarly in explaining salary variations, machine learning approaches (Supervised and Unsupervised) show promise in capturing aspects of labor market attachment that are particularly relevant to overall economic growth.

6 Robustness Analysis

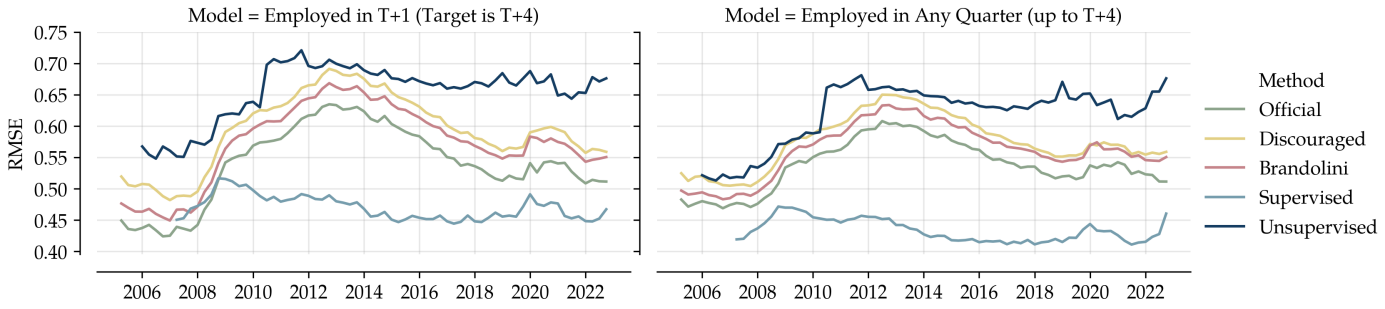
Unsupervised Method Robustness. LDA inference relies on probabilistic methods such as Gibbs sampling or variational inference, which involve stochastic processes. Changing the random seed affects the initial assignment of variables to LMA categories, leading to slight variations in the final LMA distributions. While the overarching structure of LMA may remain stable, differences in variable-LMA assignments and LMA prevalence across individuals can emerge due to the randomness inherent in the inference procedure. For that reason, all of our calculations for the unsupervised methodology are a 10 different, random seed average. This ensures robustness and discards any source of variation or noise in the inferred unsupervised LMA distribution.

In the LDA model, the hyperparameters the α and β play a fundamental role in shaping the inferred structure. Adjusting these parameters influences the distribution of the LMA across individuals and the allocation of variables within LMA states, leading to different modeling outcomes. This parameter governs the distribution of LMA across individuals. A higher α value promotes a more uniform LMA distribution, meaning that individuals are more likely to be associated with multiple LMA categories in relatively balanced proportions. In contrast, a lower β induces sparsity, leading to individuals being predominantly associated with only a few LMA categories. This sparsity can be useful when individuals are expected to belong primarily to distinct labor market groups rather than exhibiting a mix of multiple attachment patterns. In Figure 19 we observe how the reconstructed unemployment rate with different hyperparameters lead to no changes in the predicted unemployment rate variation and level.

Supervised Method Robustness Equation (8) could me minimized using another estimator,

such as the Random Forest algorithm. Although this algorithm is not directly interpretable, it has shown superb performance in many applications and is able to leverage non-linearities within the data (Hastie et al., 2009). Figure 18 in the Appendix shows the quarterly ROC AUC for such method in estimating the supervised LMA. There are no noticeable differences between Figure 15 and Figure 18, meaning that the use of one vs. another wouldn't make a big difference in the resulting LMA. In fact, we prefer the former, (the Logistic regression classifier with L1 penalty) since we can disentangle the drivers (fitted parameters) of the model. Also, we wonder whether the use of an alternative dependent variable during the training process of the supervised method would deliver better results. Instead of using the following year as the target or variable we're trying to predict to estimate the LMA, we just use whether an individual was employed in $t + 4$. The fitting metrics can be seen in the Appendix, Figure 16. Results are not altered in terms of statistical modeling.

Figure 13: Predicting Employment: RMSE



Notes: Sample restricted to non-retired and non-employed. The *discouraged* classification adds w.r.t the *official* classification the following: OLF people not looking for a job, available to work and willing to work. The *Brandolini* classification restricts w.r.t the *discouraged* classification to being looking for a job in the previous period. Left graph shows the RMSE but with the Supervised methodology using the $t + 4$ target as explained above. Right graph shows both unsupervised and supervised baseline methodologies; but instead of predicting the next quarter's employment, we're comparing at the individual level whether that individual was employed in any quarter during the following year (next four quarters). The standard errors are three orders of magnitude smaller than the values shown in this Figure; therefore, they are visually indistinguishable, suggesting high precision in the estimates across all methodologies.

However, do our results change in terms of predicting flows into employment? In Figure 13 we see that results are not altered. Left graph shows the RMSE but with the Supervised methodology using the $t + 4$ target as explained above. Right graph shows both unsupervised and supervised baseline methodologies; but instead of predicting the next quarter's employment, we're comparing at the individual level whether that individual was employed

in any quarter during the following year (next four quarters).

7 Results for other countries

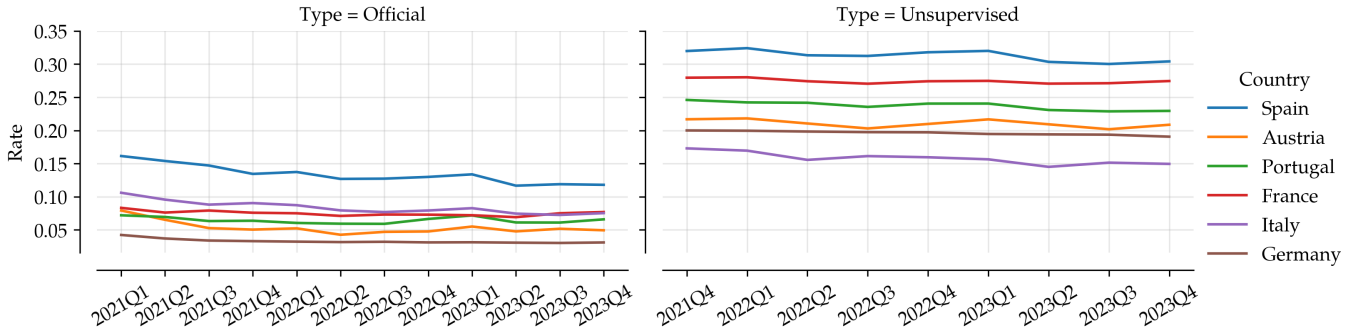
The methods presented in this paper involve strict data requirements, which may limit the implementation for other countries. First, the unsupervised method needs access to micro-data of non-employed individuals, ideally covering a broad set of socio-economic variables. And further, the supervised procedure requires following the individuals over time. We have created a repository in https://github.com/nicoforteza/labor-market-attachment-machine-learning with all the code that implements our methodology in the hope that it will serve as a guide for future practitioners.

We have computed both methods using LFS data for several countries, provided by Eurostat. Unfortunately, many of the needed variables are not available in the dataset until 2021. Further, an individual identification number, which would be needed to track individuals over time, is not available for researchers. As a consequence, we can only compute our unsupervised method, and only from 2021 on.

With these data limitations, we have estimated our unsupervised measure of attachment for some European countries, and computed aggregate attachment ratios equivalent to unemployment rates, as described in previous section. The results are presented in Figure 14. The limited time span prevents us from analyzing any temporal or cyclical patterns. However, the comparison, both across countries and through methodologies, provides useful information. The first apparent result is that for all countries, the unsupervised estimation yields higher attachment in comparison with unemployment. Hence, this is not a particular feature present in Spain, meaning that for all countries analyzed, unemployed are very close to a broader set of people who may also be considered attached, based on the similarity of other variables. Second result is that Spain is on top of the ranking in both measures, but the distance to other countries in the unsupervised estimation is not as high as in the case of unemployment. Hence, our methodology does not change the fact that Spain has a high share of attached people without a job, but our results suggest that differences with respect

to other countries could be smaller than those observed in unemployment.

Figure 14: Unsupervised LMA for European Countries



Notes: LMA is computed using European LFS microdata for Austria, France, Germany, Italy and Portugal. Y-axis shows the attachment for two methodologies. For the Official methodology, attachment is equal to Unemployment. For the Unsupervised methodology, attachment is equal to the fitted LMA probability. Note that in order to compute such probability, we need at least 1 year of historical data.

8 Concluding Remarks

In this paper we propose a novel machine learning methodology that takes into account individual heterogeneity to estimate labor market attachment. Applied over the whole non-employed population in Spanish survey data, such estimation has a continuous multi-modal probability distribution, which means that our methodology is able to recover the degree of attachment to the labor market of an individual. This contrasts with the branch of the literature of labor force composition, that uses rule-based estimators to deliver categorical classifications, which potentially can bias attachment measurement. In fact, we prove that such Official (and other basic alternative methodologies of measuring unemployment) are worse off in predicting flows to employment and the real economic cycle. Estimation and validation of the methods were done without the use of any future information, which implies that these alternate measures of LMA can be computed in real time. Finally, our work illustrates the importance of a correct measure of LMA, beyond the usual distinction of unemployed vs. inactive. We have focused on methods dividing population in two groups for comparison purposes. But, in this regard, the analisis of the optimal number of clusters (beyond two) in which we can divide non-employed people in terms of labor market attachment seems a promising avenue of future research.

References

Abowd, John M., and Arnold Zellner. (1985). "Estimating gross labor-force flows". *Journal of Business Economic Statistics*, 3 (3), pp. 254–283.
http://www.jstor.org/stable/1391596

Ahn, Hie Joo, and James D. Hamilton. (2022). "Measuring labor-force participation and the incidence and duration of unemployment". *Review of Economic Dynamics*, 44, pp. 1–32.
https://doi.org/https://doi.org/10.1016/j.red.2021.04.005

Barnichon, Regis, and Andrew Figura. (2016). "Declining desire to work and downward trends in unemployment and participation". *NBER Macroeconomics Annual*, 30, pp. 449–494.
https://doi.org/10.1086/685969

Blanchard, Olivier Jean, and Peter Diamond. (1992). "The flow approach to labor markets". *The American Economic Review*, 82 (2), pp. 354–359.
http://www.jstor.org/stable/2117427

Blei, David M., Andrew Y. Ng and Michael I. Jordan. (2003). "Latent dirichlet allocation". *Journal of Machine Learning Research*, 3, p. 9931022.

Brandolini, Andrea, Piero Cipollone and Eliana Viviano. (2006). "Does the Ilo Definition Capture All Unemployment?" *Journal of the European Economic Association*, 4 (1), pp. 153–179.
https://doi.org/10.1162/jeea.2006.4.1.153

Clark, Kim B., Lawrence H. Summers, Charles C. Holt, Robert E. Hall and Martin Neil Baily. (1979). "Labor market dynamics and unemployment: A reconsideration". *Brookings Papers on Economic Activity*, 1979 (1), pp. 13–72.
http://www.jstor.org/stable/2534304

Conde-Ruiz, J. Ignacio, Manu García, Luis A. Puch and Jesús Ruiz. (2019). "Calendar effects in daily aggregate employment creation and destruction in Spain". *SERIEs*, 10 (1), pp. 25–63.
https://doi.org/10.1007/s13209-019-0187-7

Cuadrado, Pilar, and Federico Tagliati. (2018). "La moderación salarial en España y en la UEM". *Boletín Económico - Banco de España*, 4/2018.
https://repositorio.bde.es/handle/123456789/8378

Darby, Michael R, John C Haltiwanger and Mark W Plant. (1986). "The ins and outs of unemployment: The ins win". Working Paper, 1997, National Bureau of Economic Research.
https://doi.org/10.3386/w1997

Devereux, Paul J., and Robert A. Hart. (2006). "Real wage cyclicality of job stayers, within-company job movers, and between-company job movers". *ILR Review*, 60 (1), pp. 105–119.
https://doi.org/10.1177/001979390606000106

Elsby, Michael W.L., Bart Hobijn and Ayegül ahin. (2015). "On the importance of the participation margin for labor market fluctuations". *Journal of Monetary Economics*, 72, pp. 64–82.
https://doi.org/https://doi.org/10.1016/j.jmoneco.2015.01.004

Feldstein, Martin, and James Poterba. (1984). "Unemployment insurance and reservation wages". *Journal of Public Economics*, 23 (1), pp. 141–167.
https://doi.org/https://doi.org/10.1016/0047-2727(84)90070-7

Feng, Shuaizhang, and Yingyao Hu. (2013). "Misclassification errors and the underestimation of the us unemployment rate". *American Economic Review*, 103 (2), pp. 1054–70.
https://doi.org/10.1257/aer.103.2.1054

Flinn, Christopher J., and James J. Heckman. (1983). "Are unemployment and out of the labor force behaviorally distinct labor force states?" *Journal of Labor Economics*, 1 (1), pp. 28–42.
https://doi.org/10.1086/298002

Font, Paulino, Mario Izquierdo and Sergio Puente. (2015). "Real wage responsiveness to unemployment in spain: asymmetries along the business cycle". *IZA Journal of European Labor Studies*, 4 (13).
https://doi.org/10.1186/s40174-015-0038-x

Fuchs, Johann, and Enzo Weber. (2017). "Long-term unemployment and labour force participation: a decomposition of unemployment to test for the discouragement and added worker hypotheses". *Applied Economics*, 49 (60), pp. 5971–5982.
https://doi.org/10.1080/00036846.2017.1368991

Garrido, Luis, and Luis Toharia. (2004). "What does it take to be (counted as) unemployed? the case of spain". *Labour Economics*, 11 (4), pp. 507–523. European Association of Labour Economists 15th Annual Conference, Universidad Pablo de Olavide, Seville, 18-21 September 2003.
https://doi.org/https://doi.org/10.1016/j.labeco.2004.01.007

Gertler, Mark, Christopher Huckfeldt and Antonella Trigari. (2020). "Unemployment Fluctuations, Match Quality, and the Wage Cyclicality of New Hires". *The Review of Economic Studies*, 87 (4), pp. 1876–1914.
https://doi.org/10.1093/restud/rdaa004

Hall, Robert E., R. A. Gordon and Charles Holt. (1970). "Why is the unemployment rate so high at full employment?" *Brookings Papers on Economic Activity*, 1970 (3), pp. 369–410.
http://www.jstor.org/stable/2534138

Hall, Robert E., and Sam Schulhofer-Wohl. (2018). "Measuring job-finding rates and matching efficiency with heterogeneous job-seekers". *American Economic Journal: Macroeconomics*, 10 (1), pp. 1–32.
https://doi.org/10.1257/mac.20170061

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani and Jerome Friedman. (2009). "Random forests". *The elements of statistical learning: Data mining, inference, and prediction*, pp. 587–604.

Ho, Kong Weng, and Randy Tan. (2008). "Nonmonotonic relationship between human capital and unemployment: an exploratory study with empirical evidence on singapore". *Applied Economics Letters*, 15 (15), pp. 1177–1185.
https://doi.org/10.1080/13504850500461399

Jones, Stephen R. G., and W. Craig Riddell. (1999). "The measurement of unemployment: An empirical approach". *Econometrica*, 67 (1), pp. 147–161.
http://www.jstor.org/stable/2999498

Jones, Stephen R. G, and W. Craig Riddell. (2006). "Unemployment and Nonemployment: Heterogeneities in Labor Market States". *The Review of Economics and Statistics*, 88 (2), pp. 314–323.
https://doi.org/10.1162/rest.88.2.314

Jones, Stephen R. G., and W. Craig Riddell. (2019). "Unemployment, marginal attachment, and labor force participation in canada and the united states". *Journal of Labor Economics*, 37 (S2), pp. S399–S441.
https://doi.org/10.1086/703399

Kroft, Kory, Fabian Lange, Matthew J. Notowidigdo and Lawrence F. Katz. (2016). "Long-term unemployment and the great recession: The role of composition, duration dependence, and nonparticipation". *Journal of Labor Economics*, 34 (S1), pp. S7–S54.
https://doi.org/10.1086/682390

Krueger, Alan B., Judd Cramer and David Cho. (2014). "Are the long-term unemployed on the margins of the labor market?" *Brookings Papers on Economic Activity*, 2014 (1), pp. 229–299.
https://doi.org/https://doi.org/10.1353/eca.2014.0004

Lange, Fabian, and Marianna Kudlyak. (2014). "Measuring Heterogeneity in Job Finding Rates among the Nonemployed Using Labor Force Status Histories". *IZA Discussion Paper*, (8663).
https://docs.iza.org/dp8663.pdf

Lucas, Robert E., and Leonard A. Rapping. (1969). "Real wages, employment, and inflation". *Journal of Political Economy*, 77 (5), pp. 721–754.
http://www.jstor.org/stable/1829964

Moffat, John, and Hong Il Yoo. (2015). "Who are the unemployed? evidence from the united

kingdom". *Economics Letters*, 132, pp. 61–64.
https://doi.org/https://doi.org/10.1016/j.econlet.2015.04.017

Poterba, James M., and Lawrence H. Summers. (1986). "Reporting errors and labor market dynamics". *Econometrica*, 54 (6), pp. 1319–1338.
http://www.jstor.org/stable/1914301

Rebollo-Sanz, Yolanda. (2012). "Unemployment insurance and job turnover in spain". *Labour Economics*, 19 (3), pp. 403–426.
https://doi.org/https://doi.org/10.1016/j.labeco.2012.02.008

Shibata, Ippei. (2022). "Reassessing classification errors in the analysis of labor market dynamics". *Labour Economics*, 78, p. 102252.
https://doi.org/https://doi.org/10.1016/j.labeco.2022.102252

Shimer, Robert. (2012). "Reassessing the ins and outs of unemployment". *Review of Economic Dynamics*, 15 (2), pp. 127–148.
https://doi.org/https://doi.org/10.1016/j.red.2012.02.001

Stephan Klasen, Janneke Pieters, TU THI NGOC Le, and Manuel Santos Silva. (2021). "What drives female labour force participation? comparable micro-level evidence from eight developing and emerging economies". *The Journal of Development Studies*, 57 (3), pp. 417–442.
https://doi.org/10.1080/00220388.2020.1790533

A Variables

All variables are categorical. Specifically, for each category within a variable, we create a binary dummy variable. For example, for the *age* variable there are 5 binary variables in the dataset.

LFS Variables

Interactions of variables: age and time since last employment, age and education, age and registration dummies in the public employment office, education and time since last employment, education and registration dummies in the public employment office.

Sociodemographics

- **Age**: 16 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64.
- **Student**: Yes.
- **Nationality**: Spanish.
- **Sex**: Male, woman.
- **Education**: Primary education or incomplete primary education, secondary education (first stage), secondary education (second stage), higher education.
- **Currently studying**: Has formal education or is a vacationing student, no formal education in the last 4 weeks.

Employment

- **Search of Employment Reasons**: Has sought employment or made arrangements to become self-employed in the last 4 weeks, because no suitable employment is available, because he/she is affected by an employment regulation, due to own illness or disability, for caring of sick, disabled, elderly or children or adults, for other family reasons, due to have being studying, due to being retired , for other reasons or don't know.

- **Time Seeking for Employment**: Less than 1 month, 1 to < 3 months, 3 to < 6 months, 6 months to < 1 year, 1 year to < 1 year and a half , from 1 year and a half to < 2 years , from 2 to < 4 years, 4 or more years.

- **Found Employment**: Has found employment: missing, has found and will start working within 3 months or less, has found and will be on board within 3 months or more, has not found but wish to find employment, has not found and would not wish to find employment.

- **Employment Availability**: Availability to work within 2 weeks is NaN, it is available to start work within two weeks., not available to start work within 2 weeks due to education, unavailable to work due to own illness or incapacity, not available due to care responsibilities for children or other family members, not available for other reasons.

- **Time since last employment**: Less than 1 month since leaving last job, quintile 1 of time in months since leaving last employment, quintile 2 of time in months since leaving last employment, quintile 3 of time in months since leaving last employment, quintile 4 of time in months since leaving last employment, quintile 5 of time in months since leaving last employment, has not previously worked.

- **Self-perceived Inactivity Status**: Student (even if on vacation), received a retirement pension or a pre-retirement income, engaged in housework, permanently disabled, receiving a pension other than retirement (or early retirement)., performing unpaid social work, charitable activities..., other situations.

- **Public Employment Office**: Registered and receiving unemplyment benefit or subsidy, registered without receiving unemployment benefit or subsidy, not registered but receiving unemployment benefits or subsidies, not registered and not receiving unemployment benefits or subsidies, dk/na.

LAST OCCUPATION CHARACTERISTICS

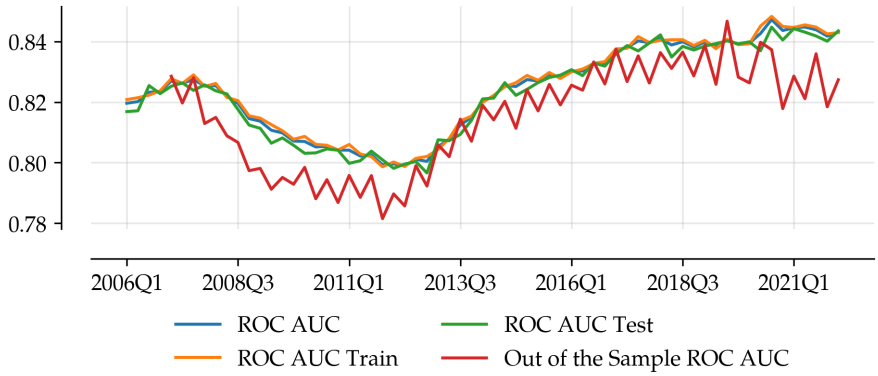
- **Last job occupation**: Military, directors and managers, technicians and scientific and

intellectual professionals, technicians, support professionals, accountants, clerical and other office jobs, food service, personal, protection and sales workers, skilled workers in the primary sector, artisans and workers in manufacturing and construction (excludes operators), plant and machine operators and assemblers, elementary occupations.

- **Last job economic sector**: Agriculture, livestock, forestry and fishing, food, textile, leather, wood and paper industries, extractive industries, petroleum refining, chemical industry, pharmaceuticals, rubber and plastics industry, electricity, gas, steam and air conditioning supply, water supply, waste management. metallurgy, construction of machinery, electrical equipment and transportation material. industrial installation and repair, construction, wholesale and retail trade and its installations and repairs. repair of automobiles, hotel and catering trade, transportation and warehousing. information and communications , financial intermediation, insurance, real estate activities, professional, scientific, administrative and other services., public administration, education and health activities, other services.

- **Professional Status of Last Employment**: Employer with employees, self-employed or entrepreneur without employees , member of a cooperative , helping in the family business or company , salaried employee in the public sector, private sector salaried employee, other .

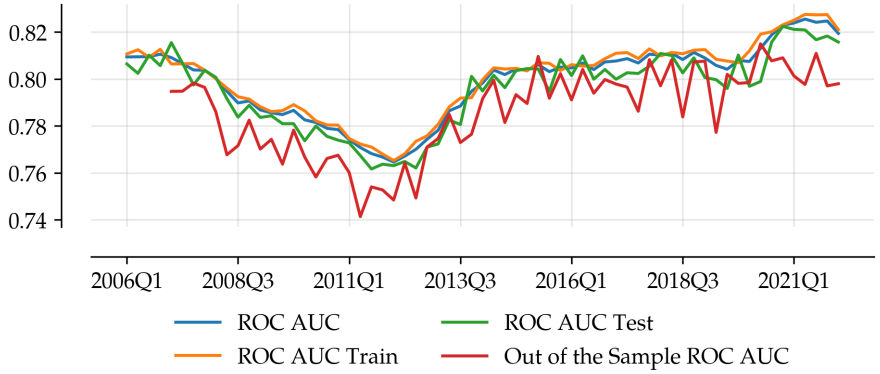
Figure 15: Supervised Method Model Validation



(a) *Notes:* The ROC AUC (Receiver Operating Characteristic Area Under the Curve) measures how well the model distinguishes between positive (any quarter employed between $t+1$ and $t+4$) and negative classes (not employed any of the next fourth quarters) across all possible classification thresholds. It considers both True Positive Rate (TPR) (or recall) and False Positive Rate (FPR). The red line shows the comparison between the prediction of the model, and the others the metric with the observed in-sample data (training and test phase)

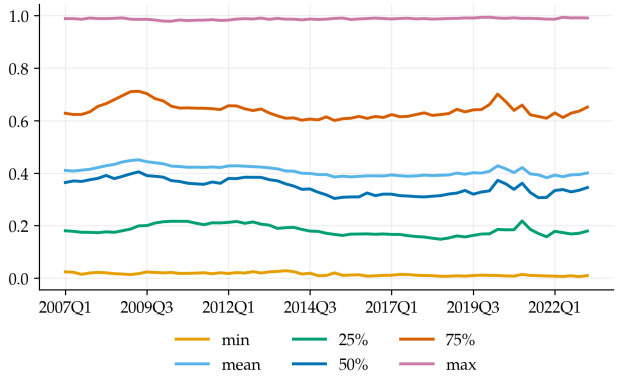
B Figures

Figure 16: Supervised Method Model Validation (T+4)



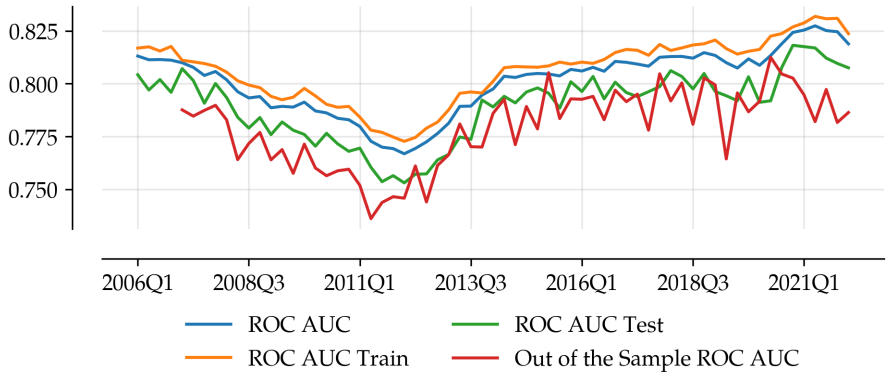
(a) *Notes:* The ROC AUC (Receiver Operating Characteristic Area Under the Curve) measures how well the model distinguishes between positive (employed in $t+4$) and negative classes (not employed in $t+4$) across all possible classification thresholds. It considers both True Positive Rate (TPR) (or recall) and False Positive Rate (FPR). The red line shows the comparison between the prediction of the model, and the others the metric with the observed in-sample data (training and test phase)

Figure 17: Supervised Method Descriptive Statistics



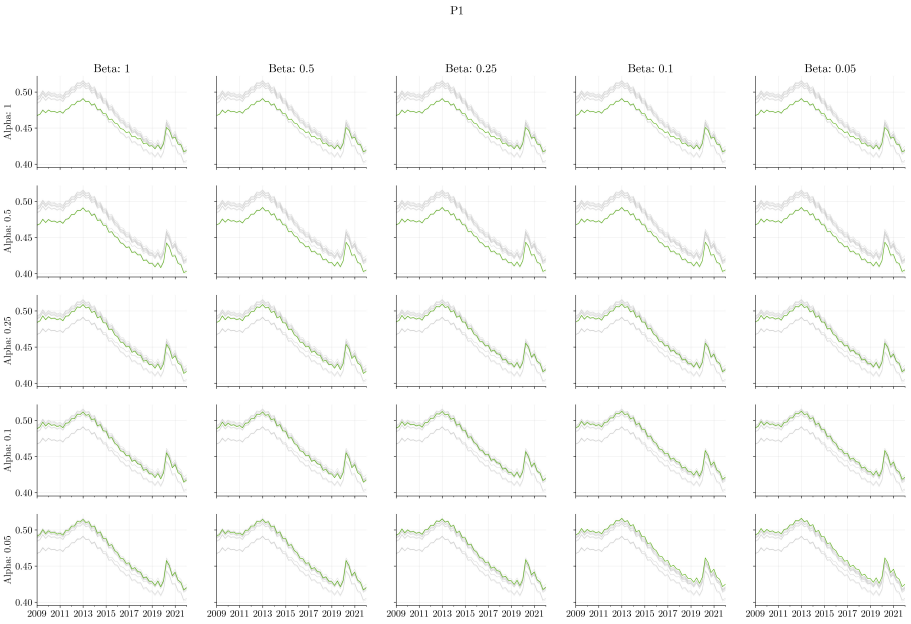
(a) *Notes:* We show the main descriptive statistics of Figure 3b. The median and mean show practically the same value (0.4). We use such threshold to separate the population in attached ($p > 0.4$) and non-attached to the labor market ($p <= 0.4$)

Figure 18: Random Forest Validation



(a) *Notes:* The ROC AUC (Receiver Operating Characteristic Area Under the Curve) measures how well the model distinguishes between positive (employed the following year) and negative classes (not employed the following year) across all possible classification thresholds. It considers both True Positive Rate (TPR) (or recall) and False Positive Rate (FPR). The red line shows the comparison between the prediction of the model, and the others the metric with the observed in-sample data (training and test phase)

Figure 19: LDA Results with Varying Hyperparameters'



(a) *Notes:* A higher α value promotes a more uniform LMA distribution, meaning that individuals are more likely to be associated with multiple LMA categories in relatively balanced proportions. In contrast, a lower β induces sparsity, leading to individuals being predominantly associated with only a few LMA categories.

C Tables

Table 1: Labor Market Attachment vs. Economic Outcomes (with and without COVID)

	Total Salaries					Salaries (market)					Real GDP rate				
Official	-0.256***					-0.272***					-0.475***				
	(0.035)					(0.036)					(0.068)				
	-0.237***					-0.238***					-0.472***				
	(0.034)					(0.031)					(0.070)				
Discouraged		-0.246***					-0.262***					-0.471***			
		(0.034)					(0.034)					(0.064)			
		-0.226***					-0.227***					-0.463***			
		(0.033)					(0.029)					(0.065)			
Brandolini			-0.251***					-0.267***					-0.468***		
			(0.034)					(0.034)					(0.066)		
			-0.231***					-0.232***					-0.462***		
			(0.034)					(0.030)					(0.068)		
Supervised				-0.093*					-0.111**					-0.642***	
				(0.051)					(0.052)					(0.055)	
				-0.069					-0.075					-0.629***	
				(0.049)					(0.045)					(0.056)	
Unsupervised					-0.382***					-0.393***					-0.771***
					(0.056)					(0.059)					(0.101)
					-0.350***					-0.340***					-0.751***
					(0.054)					(0.050)					(0.104)
Observations	65	65	65	65	65	65	65	65	65	65	65	65	65	65	65
	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63
R^2	0.731	0.732	0.735	0.524	0.714	0.777	0.779	0.781	0.595	0.749	0.773	0.786	0.778	0.873	0.791
	0.715	0.713	0.716	0.504	0.699	0.735	0.735	0.737	0.488	0.701	0.546	0.563	0.549	0.743	0.573

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*