*LABOUR*

ORIGINAL ARTICLE OPEN ACCESS

# Analyzing Non-Random Selectivity in Online Job Advertisements Using Eurostat Benchmark Data and Generalized Sample Selection Models: An Application to EU Regional Labor Markets

Pietro Giorgio Lovaglio | Mario Mezzanzanica

Department of Statistics and Quantitative Methods, University of Milano-Bicocca and CRISP, University of Milano-Bicocca, Milan, Italy

**Correspondence:** Pietro Giorgio Lovaglio (piergiorgio.lovaglio@unimib.it)

**Accepted:** 10 January 2026

**Keywords:** Eurostat benchmark data | Labor Force Survey | missingness mechanism | multilevel modeling | online job advertisements | post-stratification | representativeness | sample selection models

## ABSTRACT

The present paper provides an overall framework to afford the problem of non-representativeness and non-random selectivity arising from online job ads data, using Generalized sample selection models and Eurostat benchmark data. We jointly model the outcome intensity (number of online job ads in observed profiles, whose levels are defined by auxiliary variables) and the probability of endogenous selection (likelihood that online job ads are not missing in a given profile), allowing us to model the missing data mechanism without the need of a priori justification of missingness at random, as generally supposed by multilevel regression and post-stratification, a popular benchmark technique in this field. Moreover, we offer new post-stratification strategies to calibrate the unconditional predictions on benchmark/reference samples. We use data from the Cedefop's Skill Ovate platform collecting online job advertisements for all EU regions in 2022 and an Italian web-platform during 2013Q2-2018Q2, whereas as reference samples, aggregated LFS recent job starters and LFS new hires from microdata that represent reasonable lower bounds for job advertisements. Online job ads present a strong overrepresentation with respect to benchmark data (+40% with respect to LFS recent job starters and +400% over new hires from LFS microdata), whereas generalized sample selection models reduced this bias by half, unlike Multilevel post-stratification and other univariate approaches, which furthermore resulted in bias.

**JEL Classification:** C13, J21, J23

## 1 | Introduction

Vacancies are a crucial variable for policymakers (Cedefop 2019). In deciding the monetary policy stance, Central Banks look at the vacancy rate as a key indicator for labor market tightness.

The negative relationship between the unemployment rate and the vacancy rate, known as the Beveridge curve, is used by economists to gain insights about the relevant mechanisms underlying the functioning of the labor market (e.g., Askitas and Zimmermann 2009; Bokányi et al. 2017; Caperna et al. 2020; Fondeur and Karamé 2013). As an example, an outward shift of the Beveridge curve observed in the United States and in Europe following the great financial crises has been interpreted as a deterioration in the hiring/matching process of the economy.

In this perspective, an innovative and very promising source of data is represented by online job advertisements (Choi and

Varian 2012; Schmidt and Vosen 2013; Askitas and Zimmermann 2009, 2015; Fondeur and Karamé 2013; de Pedraza et al. 2019; Lovaglio et al. 2020; Štefánik et al. 2022).

Specifically, online job advertisements (OJAs) have emerged as an important source of information for understanding labor market trends in recent years, due to their high levels of granularity and timely information not typically addressed by official sources (Cedefop 2023; Japec et al. 2015; Couper 2013; National Science Foundation 2018).

Recently, OJA data have been explored in relation to economic theories, such as the Beveridge curve (Turrell et al. 2018), the Phillips curve (Faryna et al. 2022), the role of technological change (Hershbein and Kahn 2018), or that of skills in Mincerian equations (Deming and Kahn 2018), as indicators of new employment trends (Lovaglio 2022), or to measure skill mismatches and labor shortages (Cedefop 2023, 2024).

Despite the high detail and granularity of such data sources, their limitation, as typically occurs with web data, is that OJA may not be representative of the whole population of vacancies or regarding the structure of employment, as emerges from official statistics.

Recently, the American Association for Public Opinion Research (AAPOR) Task Force has proposed a "methodological agenda" for Big Data, a very useful document to statisticians who want to use Big Data or contribute to the Big Data debate.

In this end, the AAPOR report (Japec et al. 2015) opened to the scientific community two major challenges, such as to understand the nonresponse problem and to use and develop better statistical tools for the validity of big data in order to integrate big data and traditional surveys, as suggested by other leading authors (Couper 2013; Tam and Clarke 2015).

Specifically, within the Big Data Total Error framework cited in AAPOR report, three main sources (row, column, and cell errors) were mentioned as its components and among them the most problematic issues rely on the first: the problem of missing/incomplete data (Kreuter and Peng 2014) and particularly the non-random selectivity of population members. The issue of representativeness is linked to the presence of a non-random (not ignorable) missing data process (Kreuter and Peng 2014), which arises from non-random selectivity or the self-selection of population members in the sample (Fan et al. 2014; Fan and Liao 2012; Gelman 2007; Wu and Carroll 1988).

In our specific context, OJA data are prone to *selectivity*, *since* platforms' collection procedures are not set up for statistical purposes, and observed samples of online vacancies are likely to be affected by a non-random mechanism (not all online job advertisements are collected, not all websites are covered, advertisements non-conveyed through the web, sector and/or occupations which are [under/over-represented]). As a result, selectivity causes coverage and non-response (or missingness) that introduce potential bias in estimates based on OJA data.

The lack of representativeness in online job advertisements has strong economic relevance.

Selective coverage of vacancies may distort measures of labor market tightness, potentially biasing the inferred relationship between unemployment and vacancies in the Beveridge curve, which is widely used to assess matching efficiency and cyclical dynamics (Diamond and Şahin 2014; Marinescu and Wolthoff 2020). Under- or over-representation of particular sectors or occupations could therefore lead to misleading conclusions about structural versus cyclical shifts in labor demand, with direct implications for labor market policy and monitoring.

For instance, if online vacancies are disproportionately high because they overrepresent high-skilled occupations while low-skilled or informal sector vacancies are underrepresented, the vacancy side of the Beveridge curve will be biased upward, giving the impression of stronger labor demand than actually exists (Şahin et al. 2014). Conversely, if online platforms under-sample public sector or small-firm job postings, vacancy dynamics may appear less responsive to unemployment than they are in reality, which could be misinterpreted as evidence of declining matching efficiency (Hobijn and Şahin 2013). Evidence from empirical comparisons confirms this concern: Hershbein and Kahn (2018) show that vacancy indices based on online data diverged from survey-based measures during the Great Recession, overstating the speed of recovery in labor demand.

Such distortions matter for policy, since they affect the assessment of labor market tightness and the diagnosis of whether shifts in the Beveridge curve reflect cyclical slack or structural change.

The relevance of this selectiveness is particularly acute during periods of rapid labor market adjustment, such as recessions or recoveries, when policymakers rely on vacancy measures to gauge shifts in labor demand. If OJA data overrepresent expanding sectors (e.g., ICT, finance) while underrepresenting shrinking or stagnant ones (e.g., low-wage services), aggregate vacancy indicators may signal a stronger rebound than survey-based data suggest, leading to overly optimistic assessments of labor market tightness. Similarly, cross-regional disparities in online posting intensity can distort the geographic Beveridge curve, masking localized mismatches between vacancies and unemployment (Marinescu and Rathelot 2018).

Correcting these biases through proper statistical approaches not only improves the accuracy of tightness indicators but also enhances their usefulness for policy decisions, such as evaluating the effectiveness of active labor market programs or anticipating wage pressures in specific sectors.

In this perspective, representativeness is a very general concept that can have different meanings and shades. For our purpose, we adopt the definition of Kruskal and Mosteller who state that: "in a representative sampling, every individual in a particular target population has the same probability of being sampled, thus a sample is said to be representative with respect to a variable if its relative distribution in the sample is equal to its relative distribution in the population" (Kruskal and Mosteller 1979). Therefore, a representative sample is a miniature of the population it originates from and this should be reflected in the sample being representative with respect to all the relevant variables. Those variables are generally named auxiliary variables that

define the population strata (i.e., covariates known through census or register data that are observable also for the non-respondents). Thus, they are crucial to define and measure sample representativeness in terms of auxiliary variables.

Elliott and Valliant (2017); Valliant (2019); Beręsewicz et al. (2018); Buelens et al. (2018, 2015) give a general overview of possible approaches to deal with non-probability samples including pseudo-randomization and the model-based approach (traditional and machine learning).

Particularly, within a domain level, various selectivity strategies at *unit level* such as reweighting models (model-free and model assisted calibration, propensity weighting, two-step weighting) that require sampling frames, initial sampling weights (to correct coverage) as well as a list of respondents from possible respondents (to correct non-response) are not feasible. The same applies to the individual modeling approach that estimates a model which will then be used with a representative sample to make estimations for the target population.

Another possible strategy to deal with representativeness at domain level is a modeling approach based on detailed variables collected on the sample that are used to predict values for the non-sample units.

This rests on the superpopulation modeling approach (Elliott and Valliant 2017) to assess nonprobability samples, where a statistical model is fitted for an outcome variable from the sample and used to project the sample to the full population. If this model has the same form as for non-sample units (unobserved/missing observations), then the model fitted from the sample can be used to predict values for the non-sample. This rests on independence of sample and non-sample data, conditional on the covariates. If this does not hold, the observed portion of the population differs from the unobserved portion and this constitutes a standard case of non-random sample truncation (J. Heckman 1976), or non-random selectivity (Fan et al. 2014; Fan and Liao 2012; Gelman 2007; Wu and Carroll 1988), or the presence of selective forces (Kruskal and Mosteller 1979), all synonyms of a Missing Not at Random (MNAR) situation.

Among sophisticated statistical models for converting these inherently biased samples into unbiased estimates, the so-called multilevel regression and post-stratification (MRP) strategy (Gelman 2007; Gelman et al. 2016; Wang et al. 2014), exploiting the hierarchical data structure (e.g., individuals within regions within states), was used for integrating Web data and survey/population data.

MRP is a two-step strategy, involving, firstly a multilevel equation to model an outcome variable as a function of fixed covariates and random effects, using individual data, to obtain predicted values aggregated at a population frame, whose rows or profiles are defined a priori by cross tabulation of auxiliary variables. The subsequent stage involves post-stratification, using population frequencies in each profile (derived from census or register data), to produce estimates calibrated at a population level. Hence, adjustments due to non-representativeness are performed ex post, by calibrating the sample estimates with population weights. In this setting, post-stratification has become a popular method for correcting for known differences between sample and target populations (Little 1993).

MRP has been used to obtain accurate small-area subgroup estimates, such as for public opinion and voter turnout in individual states and demographic subgroups (Park et al. 2006; Lax and Phillips 2009; Lax and Phillips 2013; Warshaw and Rodden 2012; Buttice and Highton 2013). In addition to the static MRP versions, Gelman et al. (2016) also propose a dynamic version of MRP to estimate representative dynamic measures of responses both over profiles and time.

However, this approach works essentially hypothesizing a Missing at Random (MAR) mechanism (Pfeffermann 2007), for example, assuming that all relevant (observed and unobserved) sources underlying the missingness mechanism have been specified in the multilevel model and are present (auxiliary variables) in the post-stratification frame (Little 2007).

Various strategies have been proposed in the literature to address MNAR mechanism in different contexts. Recommended approaches include conducting sensitivity analyses under alternative assumptions of the missingness mechanism (Little and Rubin 2002) and using sophisticated statistical methods from the econometric literature on sample selection models (SSM) à la Heckman (J. Heckman 1976; Maddala 1983; Valliant et al. 2000; Terza 1998). These methods can adjust for biases that may arise from a possible MNAR mechanism.

Classical SSM, however, unlike Multilevel models, typically work in cross-sectional contexts, using normally distributed outcomes and do not accommodate hierarchical data and correlations among statistical units, typically arising from longitudinal frameworks or nested data.

In this end, as relevant and flexible alternative to SSM and Multilevel, the so-called Mixed GAM Computation Vehicle (MGCV, Wood 2004, 2011) and especially Generalized Joint Regression Models (GJRM, Marra and Radice 2013), belonging to the class of semiparametric generalized linear models, are viable alternatives under MAR and MNAR mechanisms, respectively.

The present paper addresses the issue of representativeness in online job advertisements from two perspectives: (a) through empirical comparisons with benchmark labor data (new hires from the Labor Force Survey [LFS]) and (b) by evaluating methodological strategies to mitigate bias in the observed OJA.

Empirical comparisons with benchmark data are particularly important, as they identify the sectors, occupations, regions, or countries where non-representativeness in OJA is most pronounced. This evidence highlights not only where OJA data are reliable but also where caution is needed in interpretation.

Comparing different methodologies is likewise essential to determine whether predicted OJA can be considered representative of new hires, which serve as a lower-bound estimate for the population of vacancies. In this context, the population frame of new hires is used to assess the bias of predictions relative to the benchmark. Such comparisons clarify which methods

can be reliably used to draw inferences from a selected or biased sample. In particular, sample selection models (in a GJRM perspective) produce a more representative predicted count of vacancies, as they explicitly account for nonresponse in the estimation process.

The methodology is illustrated using data from the Cedefop's Skill Ovate platform collecting official European online job advertisements for all EU regions in 2022 and a well-known Italian web platform spanning from the period 2013Q2 to 2018Q2. As reference samples we use the stocks of quarterly new hires occurred within the past 3 months, as a proxy (or upper bound) of the total number of vacancies active in a given quarter, from both quarterly LFS recent job starters and a more suitable sample, selected from LFS microdata.

To our knowledge, no previous paper has addressed the two main objectives of this study. In particular, this is the first paper to measure the bias of two OJA data sources (Italian and European) by comparing them with official LFS data. Moreover, sample selection models have not previously been applied to non-representative vacancy data as a strategy to mitigate such bias.

The paper is organized as follows: Section 2 reviews the relevant literature and approaches related to working with non-representative samples. Section 3 outlines the methodological approach. Section 4 details the data utilized and the empirical strategy, while Section 5 presents the results and also outlines future challenges. Section 6 provides the conclusion.

## 2 | Literature and Main Approaches Dealing With Non-Random Selectivity

Seminal papers by relevant authors (Elliott and Valliant 2017; Valliant 2019; Buelens et al. 2018) provide a general overview of possible approaches to handle non-probability samples or truncated samples from a population. These approaches include pseudo-randomization and model-based methods, encompassing both traditional and machine learning techniques.

In the pseudo-randomization approach, various post-stratification strategies, such as reweighting models (including model-free and model-assisted calibration, propensity weighting, and two-step weighting), require sampling weights to correct for non-coverage or the sampling of population frames used with a representative sample to make estimations for the target population. The population frame represents the number of people in the population of interest (or according to a representative survey) for each combination of demographic and geographic factors, available from census data or a representative survey (auxiliary variables).

In the literature on OJA, both the Job Vacancy Statistics (JVS, Eurostat 2019) and the LFS (Eurostat 2021) were suggested as reference samples for the representativeness analysis of OJA (Beręsewicz et al. 2018; Zilian et al. (2021); Cedefop 2023, 2024).

Recent studies adopt such post-stratification strategies with reference samples using both aggregated OJA time series data (de Pedraza et al. 2019; Lovaglio et al. 2020; Lovaglio 2022) and disaggregated OJA data (Garasto et al. 2021; Cammeraat and Squicciarini 2021; Turrell et al. 2018, 2019; Hershbein and Kahn 2018).

Another possible strategy to deal with representativeness is a model-based approach which rests on a superpopulation modeling approach used to project sample predictions to the full population (Elliott and Valliant 2017). This is based on fitting models using sampled units and finding outcome predictions also for the non-sampled population units (non-response or missing observations), based on detailed sets of covariates, collected for both sampled and unsampled units.

In a collaborative spirit, institutions such as the National Science Foundation (NSF) (2018), the American Association for Public Opinion Research in the United States (Japec et al. 2015), and Eurostat in the EU (Beręsewicz et al. 2018) have suggested that combining model-based and post-stratification strategies would be a promising approach for integrating online data with survey and population data.

One of the most popular approaches among these combined methods is known in the literature as Multilevel Regression and Post-Stratification (MRP, Gelman and Little 1997; Gelman et al. 2016; Wang et al. 2014). MRP is a two-step strategy involving, first, a classical multilevel equation to model the outcome using individual data using fixed covariates (included in the sample) and random effects, exploiting the possible hierarchical data structure (e.g., individuals within regions within states or repeated measures for individuals) and, second, post-stratification.

The starting step of MRP is the choice of a reference sample or a benchmark data (here referred to as population frame), generally derived from Census or register data. The population frame, without loss of generality, contains $K$ categorical variables of interest (auxiliary variable) and that the $k$-th has $J_k$ categories.

The chosen $K$ auxiliary variables are those the researcher retains fundamental for representativeness or to recalibrate the sample estimate of an outcome variable, such as demographic and geographic characteristics (and shares) of the population. Usually, auxiliary variables may also contain continuous variables (age), and in that case, these variables will be discretized to form categorical variables (age classes).

Hence the population can be represented by $J = \prod_{k=1}^{K} = J_k$ cells (or *profiles*) calculated by cross-classification (tabulation) of all $K$ variables leading the so-called *population frame* ($P_J$), also known as post-stratification frame.

Hence, each profile (row $j$) of the population frame is thus aggregated at some level $j$ ($j = 1 \ldots J$) and their $J$ rows (*profiles*) constitute all possible profiles of the $K$ variables in the population.

Each $j$th population profile contains $N_j$ observations ($N = \Sigma_j N_j$). For example, in the context of employment analysis, $N_j$ represents the number of working individuals in each profile.

More generally, among the $K$ covariates, the set of population profiles $P_J$ may include a time dimension—such as months,

quarters, or years—depending on the selected time horizon and level of aggregation. In the employment example, $P_J$ could represent the working population across all possible combinations of Occupation × Sector × Region × Quarter, using official classifications such as ISCO-08 (occupation), Nace rev.2 (economic activity), Nuts2 (regional level), and predefined calendar quarters.

The first step of MRP is modeling individual data in the (non-probability) sample. Supposing to analyze a binary or a continuous outcome $y_i$ for $i = 1, ..., n$ observations, where the objective is to estimate a parameter $\theta$ (the prevalence of an event, or the total of $Y$ in the population, respectively).

Let $D_I$ denote the (non-probability) sample data containing individual-level observations for $i = 1, ..., n$. In the first step, a multilevel model is fitted to these data, producing individual-level predicted outcome values ($\hat{y}_i$) (e.g., the probability of being employed).

In the second step of MRP, firstly, these individual predictions are aggregated to the same level of aggregation as the population frame—that is, to the $J$ population profiles ($\hat{y}_j = n_j^{-1} \sum_{i \in j} \hat{y}_i$ where $n_j$ is the number of observations in the $j$-th profile) and secondly, they were recalibrated accordingly with strata weights ($N_j/N$) of the population frame (representative reference sample). MRP, as traditional post-stratification strategies, adjusts outcome's estimates for differences between $N_j/N$ in the population and $n_j/n$ in the sample. Hence the post-stratified estimate of $\theta$ (total) can be expressed as $\hat{y}^{PS} = \sum_j \hat{y}_j w_j / \sum_j w_j$ where $w_j = N_j/n_j$ are the weights.

However, when dealing with OJA, individual data $D_I$ (e.g., the $n \times p$ matrix of OJA with $p$ covariates describing their characteristics) is not a useful starting point for obtain $\hat{y}_i$, since $D_I$ lacks the outcome variable.

When the aim is to model the occurrence and the amount of OJA over covariates' levels (profiles) of interest, we have to reorganize $D_I$ in the same structure of the population frame.

Thus, reorganizing the sample individual data in the $J$ profiles, as above for population, this leads the *sample frame* $S_J$, where each $j$th profile has $y_j$ observations.

In the above example, $y_j$ represents the count of employed in each profile of $S_J$, whereas in our empirical setting these values represent the counts of OJA ($y_j$) in the profiles of the non-probability sample.

The population and the sample frame are matrices of dimension $J \times K+1$, where the $k$-th column collects all $J_k$ levels of the $k$-th stratification variable and the last column contains for $P_J$ the number of observations ($N_j$) and in $S_J$ the outcome's total ($y_j$) or number of observations ($n_j$) for each $j$-th profile ($j = 1, ..., J$).

Note that, although the covariates defining $S_J$ are the same as those defining $P_J$, the sample frame does not necessarily cover all $J$ levels of the population frame—particularly when certain profiles (i.e., combinations of auxiliary variables) are not represented in the non-probability sample.

For example, an ideal population frame would include the count of OJAs for all possible combinations of Nuts2 × ISCO × Nace for a given quarter in Italy. Specifically, by crossing 20 Nuts2 regions, 12 aggregated Nace sectors, and 43 occupational ISCO categories (II digits), we obtain 10,320 potential profiles.

However, many of these profiles may be missing OJA observations in the empirical data—for instance, due to the absence of OJAs for certain combinations (e.g., Associate Professionals in the Energy sector in Calabria). This results in a sample profile $S_J$ with a large number of empty or zero values in the outcome column $y_j$.

Returning to the MRP framework, unlike the typical case in which the time variable is included among the stratification (auxiliary) variables that define the profiles ($P_J$ or $S_J$), estimation strategies that leverage temporal dependence—such as multilevel models with repeated measures or random effects—can benefit from a different approach. A useful technique is to exclude the time variable from the definition of the $j$-th profile (now denoted as $j^*$). In this way, both the sample and population frames consist of a set of $J^*$ profiles, each repeated across $T$ time periods ($t = 1, ..., T$).

In this perspective, a multilevel equation can be specified in the sample frame $S_J$ where the number of *positive* OJA in a generic $j^*$-th profile at time $t$ ($y_{tj*}$) is observed.

These new profiles (where $j = t\ j^*$) were adopted only in the phase of fitting (the $J^*$ profiles were specified as random effects), whereas in the phase of post-stratification the strategy adopts the original $J$ profiles.

Hence, multilevel models work on a truncated sample $S_J$ of non-empty profiles (OJA > 0), drawn by a more general population frame $P_J$. This is the main reason of non-representativeness and selection bias affecting strategies working on a self-selected sample.

To this end, the *selection* mechanism is represented by a Bernoulli variable $z_{tj*}$ (where $z_{tj*} = 1$ when $\text{OJA}_j > 0$ and $z_{tj*} = 0$ otherwise). In practice, $y_{tj*}$ can be observed only when $z_{tj*} = 1$, otherwise $y_{tj*}$ is missing or zero. More formally, a multilevel model with random effects for profile $j^*$ at time $t$ and gaussian errors is specified, as follow:

$$y_{tj*} = \boldsymbol{\beta}_{0j*} + \boldsymbol{\beta}' \mathbf{x}_{2tj*} + \varepsilon_{tj*} \quad z_{tj*} = 1 \tag{1}$$

$$\boldsymbol{\beta}_{0j*} = \beta_0 + \boldsymbol{\delta}' \mathbf{d}_{j*} + \gamma_{j*} \tag{2}$$

$$(\varepsilon_{tj*}, \gamma_{j*}) \sim \boldsymbol{\Phi 2}(\varepsilon_{tj*}, \gamma_{j*}) = \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix}\right] \tag{3}$$

where the intercepts $\boldsymbol{\beta}_{0j*}$ vary among $J^*$ profiles (generally in MRP also the slopes can be specified as random effects: $\boldsymbol{\beta} = \boldsymbol{\beta}_{j*}$), both in a non-random fashion (depending on fixed covariates $\mathbf{d}_{j*}$, where $\boldsymbol{\delta}$ are the corresponding parameters) and in a random fashion, depending on additional set of random terms $\gamma_{j*}$ that are assumed to be i.i.d. random variables with zero mean and constant variance $\tau^2$. Moreover, these new errors $\gamma_{j*}$ are assumed

to be independent of all regressors $\mathbf{x}_2$ and $\mathbf{d}$, and as well independent of the $\epsilon_{tj*}$ errors.

Predictions $\hat{y}_{tj*}$ for all $T$ time levels of the $j^*$ profile—aligned with predictions $\hat{y}_j$ in the $j$-th profile, which also incorporates the time dimension (i.e., predictions of OJA counts across the $J$ profiles of the sample frame)—are then weighted (post-stratified) by the total number of individuals in each $j$-group of the population frame ($N_j$). This yields the overall post-stratified (PS) estimate of the total count ($\theta$):$\hat{y}^{\mathrm{PS}} = \sum_j \hat{y}_j N_j / \sum_j N_j$.

In the previous employment example, the total number of employed individuals in the population was estimated by aggregating the non-probability sample estimates ($\hat{y}_j$), appropriately weighted using counts from the population frame (i.e., official data). In the context of OJA, where population data are unavailable, a reference sample can be used to approximate the distribution of OJA across the $J$ strata in the population.

Moreover, there are situations in which, rather than seeking an overall estimate of the total in the population, we may be interested in estimating the total of an outcome variable for each $r$-th level ($r=1, \ldots, R$) of a single auxiliary variable ($\theta_r$). Each $r$-th level is obtained by aggregating $j$ levels over $r$ ($j \in r$), such that $\hat{y}_r = \sum_{j \in r} \hat{y}_j$. The corresponding post-stratified estimate of $\theta_r$ is given by $\hat{y}_r^{\mathrm{PS}} = \sum_{j \in r} \hat{y}_j N_j / \sum_{j \in r} N_j$. As example $\hat{y}_r^{\mathrm{PS}}$ represents the post-stratified proportion of employed individuals among residents in region $r$, adjusted to reflect the true distribution of employment in the overall population of that region—regardless of any over- or under-representation of individuals from region $r$ in the (non-probability) sample.

Notice that $\hat{y}_r^{\mathrm{PS}}$ have an equivalent formulation ($\hat{y}_r^{\mathrm{PS}} = \sum_{j \in r} \hat{y}_j W_j / \sum_{j \in r} \mathrm{W}_j$), written as a function of weights $W_j = n_j w_i$, where $w_i$ are the individual units' "survey weights," available in representative reference surveys (calculated as $w_i = N_i/n_i$), in order to weight the $n_i$ observations in the $i$-th row of the sample ($n_i$ is generally equals to 1) with their corresponding counts in the population ($N_i$). In our case, since the population data are aggregated at $J$ levels, $w_i$ equals to $w_j = \sum_{i \in j} w_i$, thus leading $W_j = n_j w_j = n_j(N_j/n_j) = N_j$ thus returning the original formulation of the post-stratified estimate of $\theta_r$.

Hence, post-stratification produces sample-aggregated estimates calibrated at a population level, as would be done using classical design-based weighting (Norrander 2007): individual estimates are corrected from known differences between sample and population percentages (Little 1993).

MRP has been used to obtain accurate small-area subgroup estimates, such as for public opinion and voter turnout in individual states and demographic subgroups (Park et al. 2006; Lax and Phillips 2009; Lax and Phillips 2013; Warshaw and Rodden 2012; Buttice and Highton 2013; Wang et al. 2014). Recent proposals extend MRP to machine learning techniques (Broniecki et al. 2021).

## 2.1 | MRP Limits and Extensions

Since Multilevel models incorporate random effects, they introduce an additional source of variability ($\gamma_{j*}$) that contributes to the error term of the outcome equation, also introducing additional assumptions, including the mutual independence of errors and the independence of each type of error with respect to covariates (exogeneity).

The broad range of these assumptions, many of which are generally not testable, renders multilevel models challenging to implement in various empirical contexts. The strong exogeneity hypothesis explains why many econometricians prefer alternative estimators, such as fixed effects.

This rigidity is further compounded in possible situations of non-random selectivity, where the intricate nature of the selection mechanism complicates the analysis even when only a single source of error is present. For these reasons, multilevel has limited use in the context of causal inference in a potential outcome framework (Little and Rubin 2002) or in impact analysis when the treatment is endogenous.

As relevant and flexible alternative to Multilevel, the so-called Mixed GAM Computation Vehicle (MGCV, Wood 2004, 2011), belonging to the class of semiparametric generalized linear models, extends Multilevel in two ways. First, it allows versus nonparametric specification (such as smoothing splines, loess, etc.) for covariates in the linear outcome equation. This flexibility in specifying regressor effects is particularly useful in the context of sample selection models, since parameter estimates are inconsistent when the true relationship between covariates and outcome is mismodeled (Chib et al. 2009; Marra and Radice 2011).

Second, and most importantly, it specifies the same effects that drive the hierarchy in multilevel (profiles $J^*$), not as random effects (thus avoiding new errors in the outcome equation) but as penalized fixed effects whose coefficients were estimated under some penalization constraints. Since generally the profiles are very numerous, the use of penalized coefficients is justified since we suspect that some coefficients may be weakly or not identified or under-represented in the data.

Moreover, specifying penalized coefficients for fixed effects is the essential methodological trick for estimating these coefficients as random effects (Robertson 1955; Ruppert et al. 2003; Wood 2004), exactly similar to the way they are estimated in multilevel models (known as Empirical Bayes or BLUP), but specified as random terms. The presence of fixed effects and possible non-linear (regression splines or loess) specification for covariates, MGCV generally presents excellent goodness of fit.

Equation (4) shows a general MGCV model with gaussian errors and all covariates in a linear specification:

$$y_{tj*} = \boldsymbol{\beta}_{0j*} + \boldsymbol{\beta}' \mathbf{x}_{2tj*} + \quad \boldsymbol{p}_{j*}{}' \boldsymbol{\nu} * + \quad \epsilon_{tj*} \quad z_{tj*} = 1 \tag{4}$$

where $\boldsymbol{p}_{j*} = (p_{1*}, \ldots, p_{j*}, \ldots, p_{J*})$ is a $(J \times 1)$ vector of dummy variables representing the $J^*$ profiles, with associated parameter

column vector $\boldsymbol{\nu}_* = (\nu_{1*}, ..., \nu_{j*}, ..., \nu_{J*})$ and errors $\epsilon_{tj*}$ follow a gaussian distribution with zero mean and variance $\sigma^2$. The role of $\boldsymbol{p}_{j*}$ (Profiles' dummies) is the same of that of second-level units in longitudinal or multilevel framework, where repeated measured on $\boldsymbol{p}_{j*}$ are the first level units.

As mentioned, to avoid that estimates of $\nu_{j*}$ (effect of $j$*-th profile on the outcome variable $y_{tj*}$) may be weakly or not identified, due to a large number of effects to be estimated, this issue can be dealt with by augmenting the objective function (log-likelihood $L$ as a function of all unknown parameters) with a positive penalty term only for these fixed effects $p_{j*}$. Following the framework of Additive Models (Wood 2004), these fixed effects can be expressed as $f(p_{j*})$, where f is an adopted smooth function applied to each dummy-profile effect ($p_{j*}$). The penalty term is expressed as the product of a smoothing parameter ($\lambda$), tuning the trade-off among bias and model's variance, and the integrated square of second derivative of $f(p_{j*})$. In practice, this penalty term measures the (second-order) roughness of the smooth terms in the model.

Using regression splines as $f(p_{j*})$, since they are linear in their model parameters $\nu_{j*}$ (e.g., $f(p_{j*})$ can be approximated as a linear combination of a set of Basis functions and regression coefficients $\nu_{j*}$), such a penalty can be expressed as a quadratic form in $\boldsymbol{\nu}_*$ as $\lambda\, \boldsymbol{\nu}_*^{\mathrm{T}} \mathbf{S} \boldsymbol{\nu}_*$, where $\mathbf{S}$ is a matrix of known coefficients. Hence, a penalized log-likelihood function $L_p = L - \frac{\lambda}{2}\, \boldsymbol{\nu}_*^{\mathrm{T}} \mathbf{S} \boldsymbol{\nu}_*$ was maximized to estimate all parameters.

Both ML and MGCV belong to model-based approaches that may be useful to deal with non-representative samples. However, apart from some exceptions (Matei 2018; Elliott and Valliant 2017; Zhang et al. 2013), model-based approaches rely on the assumption that missing data are MAR, that is, that conditioning on auxiliary variables (that define the population profiles), non-response is independent of the unobserved target variable. Specifically, it is supposed that the model (and estimated parameters) has the same form as for non-sample units, resting on the hypothesis of independence of sample and non-sample data, conditional on the covariates.

In this context, the model-based literature on superpopulation demonstrates that to minimize the risk of non-response bias, the emphasis should be more on the predictive power of the outcome variable that should include significant predictors, with less emphasis placed on assessing whether these predictors significantly affect the probability of missingness (Gelman 2007; Schonlau and Couper 2017; Wang et al. 2014; Zhang 2005; Zhang et al. 2013). From the MRP perspective, in fact, the focus is predominantly (if not exclusively) on the predictive power of the fitted multilevel model.

This aligns with the reasonable view that when predictors are significant only in the model for the probability of missingness, but are irrelevant in the outcome equation, it would be challenging to justify the MAR assumption.

In this end, since unbiased estimates require reasonable estimates within narrow slices of the sample, a suggested strategy in the MRP literature is to insert (deep) two-way or three-way interactions among covariate or auxiliary variables in the predictive model (Gelman et al. 2007, 2009). Deep interactions, apart from increasing the predictive power of the model (Gelman et al. 2010; Ghitza and Gelman 2013), can help correct for sampling problems and, particularly, nonresponse rates (Gelman 2007, 2014).

## 2.2 | Sample Selection Models

As above mentioned, recommended approaches for MNAR are sample selection models (SSM, J. Heckman 1976; Maddala 1983; Valliant et al. 2000; Terza 1998). SSMs jointly model both the outcome equation and the probability that an observation is missing or included in the sample (selection equation). The core idea is that, even after controlling for strongly significant covariates, other unobservable factors may influence both equations. Therefore, the two equations must be jointly estimated, acknowledging the potential correlation among the error terms. Ignoring this correlation can lead to biased and inconsistent parameter estimates.

Regarding the selection equation, using the latent variable representation, $z_j$* represents the attitude that an OJA in profile $j$ is selected (as a latent attitude to appear online) that may be determined by some vector of $M$ covariates ($\mathbf{x}_{1j}$), where $M \geq K$.

Although $z$* is unobservable, when it exceeds a certain (unknown, but without loss of generality, fixed at zero) threshold, in the sample frame $S_j$, we observe the number of OJA in a generic $j$-th profile at time $t$ ($y_j$). The *selection* mechanism is governed by the Bernoulli variable $z_j$ ($z_j = 1$ when $\mathrm{OJA}_j > 0$, and 0 otherwise), and it derives its properties from those of $z$*. Moreover, $y_j$ can be observed only when $z^* > 0$, otherwise ($z_j = 0$) $y_j$ is missing.

This generates a SSM composed by a system of two equations, such as:

$$z_j* = \pi_0 + \boldsymbol{\pi}' \mathbf{x}_{1j} + u_j \tag{5}$$

$$y_j = \beta_0 + \boldsymbol{\beta}' \mathbf{x}_{2j} + \varepsilon_j \text{ when } z_j = 1 \tag{6}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are covariates of two equations (that may share some variables), $\boldsymbol{\pi}$ and $\boldsymbol{\beta}$ the corresponding fixed parameters, and $\pi_0$ and $\beta_0$ the intercepts, respectively. The error terms $u_j$ and $\varepsilon_j$ are i.i.d. random variables assumed to be independent of their regressors $\mathbf{x}_1$ and $\mathbf{x}_2$ and following a bivariate Gaussian $\boldsymbol{\Phi}_2$ density, with zero means, constant variance ($\sigma^2$ for $\varepsilon_j$) and correlation coefficient $\rho$, namely:

$$(\varepsilon_j, u_j) \sim \boldsymbol{\Phi}_2(\varepsilon_j, u_j, \rho) = N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right] \tag{7}$$

In this formulation, Equation (5) represents a Probit model modeling the latent variable $z_j$*, while Equation (6) is a regression model with Gaussian errors. These models are jointly specified and should therefore be estimated simultaneously, in general by maximum likelihood. Separate estimations lead to biased and inconsistent estimates for parameters, as well as for predicted values (especially $\hat{y}_j$).

As known, for identification of this two-set of equations, $\mathbf{x}_1$ requires a set of instruments, for example, a subset of covariates in $\mathbf{x}_1$ not be in and $\mathbf{x}_2$ set. In our empirical context, webographics, such as internet penetration in societies and regional labor market, were used as instruments.

Although it is not possible to disentangle between MAR and MNAR situations with a formal test, we assess MAR/MNAR using both the significance of parameters ($\rho$) and the Likelihood Ratio Test for the hypothesis that $\rho = 0$.

A significant correlation coefficient among two equations ($\rho$) or a significant LR test would suggest significant sample selection, and thus that the missing data mechanism is *not ignorable* or MNAR, thus suggesting that two equations should be jointly estimated. Thus, univariate approaches such as multilevel or MGCV that ignore the selection equation are not appropriate.

### 2.3 | Discussion

Ideally, all the strategies discussed can be used to generate predictions of OJA, with the preferred methodology determined by which approach most closely aligns with benchmark data.

This comparison provides insights into the sectors, occupations, and regions where non-representativeness is most pronounced.

Although the comparison among methodologies is entirely empirical, strategies that address multiple sources of non-representativeness—such as nonresponse and missingness—are generally more desirable.

The primary advantage of sample selection models is that they can handle potential MNAR mechanisms, whereas methods such as ML/MRP and MGCV generally rely on the assumption of MAR.

MRP authors acknowledge that situations of MNAR might exist. To address this, they emphasize that, beyond focusing on the predictive power of covariates in the outcome equation, MRP should also prioritize the inclusion of highly relevant variables in the population frame, in order to "*potentially bring the missing-at-random model closer to realism*" (Little 2007).

In the same end, MRP authors (Lopez-Martin et al. 2022) remarked that although "MRP can mitigate potential biases in the sample, [...] it is not a substitute for a better data collection effort that tries to minimize systematic nonresponse patterns."

In general, hence the MRP strategy typically works when missingness is completely at random or MAR, assuming that all relevant auxiliary variables underlying the missingness mechanism have been specified in the multilevel model (Pfeffermann 2007).

Quite surprisingly, sample selection models have never been used to address non-representative vacancy data. In this end, we remark that in MRP the limitation of nonresponses is completely addressed by increasing the predictive power of the model (e.g., with deep interactions), without explicitly modeling nonresponse (selection model in our nomenclature), unlike our proposed approach that not only specifies a model for nonresponse but also models it in a bivariate joint model with outcome intensity.

Specifically, the outcome equation of SSM models the unconditional expected values $E(y_j \mid \mathbf{x}_{2j})$, thus predicting the outcome of interest for both selected (profiles with $OJA_j > 0$) and unselected observations (profiles with missing $OJA_j$).

In this context, modeling nonresponse is a suitable approach, as it allows predicted OJA to better approximate a representative sample of vacancies, enabling more accurate estimates of the distribution of predicted vacancies across regions, sectors, and occupations. These adjusted stocks should reflect the underlying economy more accurately than the non-adjusted stock of OJA.

The dissemination of bias-corrected (predicted) vacancy data, at an adequate level of granularity, may constitute a solid base providing further insights on more general labor demand analyses, such as the concentration of demanded occupations among territories or sectors or their evolution over time, or whether the demand of a given occupation varies among sectors and over time.

Another key contribution of our methodological approach is the introduction of a post-stratification strategy that recalibrates the model's predictions to benchmark totals. This strategy is particularly useful when predicted OJA are strongly biased (w.r.t. benchmark data).

Next section describes in detail the adopted methodology.

## 3 | Proposed Methodology

One complication in fitting SSM arises when modeling a non-Gaussian outcome, as in our empirical context, where the outcome consists of counts of OJA in a given profile.

For non-Gaussian outcomes, accounting for sample selection is complicated by the use of a nonlinear model to fit the data, which can result in unstable Maximum Likelihood estimators. In such cases, even the Heckman two-stage estimator is only approximate, and appropriate distribution results for the estimators are unavailable (J. J. Heckman 1979).

In the context of duration models, some authors have attempted to replicate SSM with maximum likelihood using more appropriate distributions for the intensity (duration) equation, such as exponential and Weibull distributions (Boehmke et al. 2006; Box-Steffensmeier and Jones 2004; Prieger 2002).

In the context of count data, alternatives to classic SSM are standard parametric count-data models with sample selection, such as Terza's maximum likelihood estimator (Terza 1998). Other approaches try to mitigate the sensitivity problem of non-normal errors between the two equations by using a more flexible parametric family of distributions, including non-parametric penalized maximum likelihood estimation (Wyszynski and Marra 2018) or semi-parametric approaches providing a robust two-step estimator (Zhelonkin et al. 2016).

Unfortunately, such approaches have not been extended to a multilevel context, in the sense that they do not allow specification of random effects to exploit time dependency.

In this end, fully parametric methods for estimating count-data models with normally distributed random effects were proposed (W. Greene 2001, 2012; Winkelmann 1998; Miranda 2004; Miranda and Rabe-Hesketh 2006; Boyes et al. 1989) or a two-step strategy (Kenkel and Terza 2001). Among others, the so-called Generalized linear latent and mixed models (GLLAMMs, Rabe-Hesketh et al. 2004) framework provides consistent estimators of a joint model of the outcome and selection variable that are obtained by full maximum likelihood estimation. This approach allows random effects, as well as discrete (counts) dependent variables (Rabe-Hesketh et al. 2005).

Others relevant extensions of classical SSM, known as Generalized Joint Regression Models (GJRM) are generalized sample selection models based on copulas (Lee 1983; Smith 2003; Winkelmann 2011; Marra and Radice 2013). These approaches are particularly useful and flexible as they extend sample selection models by incorporating endogenous predictors and allowing for various error distributions. This flexibility makes GJRM suitable for modeling binary, count, or highly skewed outcomes, and they provide various association structures for error terms through copula specifications (Trivedi and Zimmer 2007). Furthermore, GJRM can model the associations between errors using covariates, accommodate random effects, and include nonparametric versions of fixed covariates in the linear predictor (Ruppert et al. 2003; Wood 2004).

### 3.1 | GJRM

In this subsection, we discuss the characteristics of our methodological proposal, which consists of two phases: a GJRM modeling phase and a post-stratification phase.

First, to our purposes, instead of $S_J$ as above described, we obtain the *sparse* sample frame ($S_{J^*}$) including all possible population *profiles* (cells) derived from a (sparse) *K*-way cross-classification of the *K* covariates. The sparse approach ensures that all potential profiles are represented in the sample frame, including those levels not present in the sample data due to missing levels of some covariates in the sampled units. These *J** profiles may also cover situations of sampled units measured on a subset of the *K* covariates (case of a completely missing covariate). In these cases, additional rows to $S_{J^*}$ can be added.

Since GJRM manages repeated measures, a profile *j** of $S_{J^*}$ is defined by combinations of time-invariant covariates (e.g., Region × Sector × Occupation), each repeated *T* times. Thus, each block of *T* rows of $S_{J^*}$ represents the new *j**-th observation unit (profile) containing the outcome variable, summing individual observations in the original dataset $D_I$.

It is important to note that the number of levels of time variant and invariant covariates that define $S_{J^*}$ include all possible levels of these variables from official sources (such as Nuts2, Nace, ISCO, time), regardless of whether OJA is present in the sample data $D_I$.

On the sparse $S_{J^*}$, we can specify our GJRM model (where now the profile *j** will replace the profile *j*), composed by two equations each with profiles' dummies ($\boldsymbol{p}_{j*}$ and related parameters $\boldsymbol{\tau}_*$ and $\boldsymbol{\nu}_*$), other fixed covariates ($\mathbf{x}_1$ and $\mathbf{x}_2$) and by a general form for errors structure, such as:

$$z^*_{tj*} = \pi_0 + \boldsymbol{\pi}'\mathbf{x}_{1tj*} + \boldsymbol{p}_{j*}{}' \boldsymbol{\tau}* + u_{j*} \tag{8}$$

$$y_{tj*} = \boldsymbol{\beta}_{0j*} + \boldsymbol{\beta}'\mathbf{x}_{2tj*} + \boldsymbol{p}_{j*}\,\mathrm{T}\,\boldsymbol{\nu}* + \varepsilon_{j*} \quad z_{tj*} = 1 \tag{9}$$

$$\boldsymbol{F}(\varepsilon_{j*}, u_{j*}) = \boldsymbol{C}\big(F_1(\varepsilon_{j*}), F_2(u_{j*}),\ \rho\big) \tag{10}$$

where $\boldsymbol{F}(\varepsilon_{j*}, u_{j*})$ is the bivariate cumulative distribution function (cdf) of error terms expressed as a function $\boldsymbol{C}$, called copula, of one-dimensional margins $F_1(.)$, $F_2(.)$ and their correlation ρ.

Equation (8) models the selection mechanism in term of the latent variable $z^*_{tj*}$, underlying the observed binary variable $z_{tj*}$. Equation (10) is the crucial aspect of GJRM flexibility. Broadly speaking, a copula is a function that connect multivariate distributions to their one-dimensional margins and their association by a specific functional form $\boldsymbol{C}$, where $\boldsymbol{C}(.)$ is no more than a bivariate density. From the fundamental Sklar's (1973) theorem, if $\boldsymbol{F}$ $(y_{1j}, y_{2j})$ is a two-dimensional cdf with one-dimensional margins $F_1(y_{1j})$, $F_2(y_{2j})$, then there exists a two-dimensional copula $\boldsymbol{C}$ such that $\boldsymbol{F}(y_{1j}, y_{2j}) = \boldsymbol{C}(F_1(y_{1j}), F_2(y_{2j}); \rho)$, where $y_1$ and $y_2$ are two random variables and ρ represents an association parameter linking the correlation of two margins (Trivedi and Zimmer 2007).

Hence, the formulation $\boldsymbol{F}(\varepsilon_j, u_j) = \boldsymbol{\Phi_2}(\varepsilon_j, u_j, \rho)$ of Equation (7) is only a particular case among many possible specifications that can be expressed with copulas. In fact, using two gaussian marginals $F_1(\varepsilon_j)$, $F_2(u_j)$ for error terms and a gaussian copula $\boldsymbol{C} = \boldsymbol{\Phi}_2(\Phi^{-1}(F_1), \Phi^{-1}(F_2); \rho)$, the copula representation of $\boldsymbol{F}(\varepsilon_j, u_j)$ leads the classical standard Heckman-type model. The more flexible copula version allows non gaussian dependence among error also leading the strength and direction of the association ρ between the two marginals may vary across observations or groups of observations ($\rho_j$).

### 3.2 | Post Stratification

When modeling a count as the outcome variable ($OJA_j$), where *j* indicates the *j*-th row of the population and sample profile (e.g., all possible combinations of time invariant and time variant covariates), $N_j$ in the population frame represents the total count of individuals in the population (LFS recent starters or LFS new hires). Unlike the MRP approach, which adjusts estimates based on the percentage of sample strata in the population, our method involves a reweighting phase that recalibrates the model's estimates to population totals (LFS recent starters in row *j*).

This approach is justified since, unlike OJA that may be studied at a finer level (*j*), official labor market benchmark sources are generally provided at a more aggregated level (*s*, $s \in j$).

Specifically, *j* represents the aggregation level cross-classifying the *K* variables (e.g., Occupation × Sector × Region × Quarter) of

the sample frame where the model was fitted ($\hat{y}_j$), whereas let $s$ denotes the (less granular) aggregation level of the population frame, which is generated by cross-classifying $S \subset K$ variables (e.g., Sector × Quarter, as in the Eurostat LFS Recent Starters for each EU country). Each $s$-th profile has a total of $N_s$, which indicates the total count of the $S$-variate benchmark variable for our outcome.

In this end, our post-stratification strategy computes the weights $w_s = N_s / \hat{y}_s$, where $\hat{y}_s$ represent the sum of $\hat{y}_j$ at the $s$-th level ($\hat{y}_s = \sum_{j \in s} \hat{y}_j$), unlikely MRP that used $w_j = N_j / n_j$ as weights.

Finally, we apply these weights to the most granular estimates $\hat{y}_j$ (post-stratification) to obtain the overall post-stratified estimate of the parameter of interest θ (e.g., total amount of OJA in the population):

$$\hat{y}^{\mathrm{PS}} = \frac{\sum_{j \in s} w_s \hat{y}_j}{\sum_{j \in s} w_s} \tag{11}$$

or the post stratified estimates of a more disaggregated $r$-th level $\theta_r$, (e.g., total amount of OJA in the population by $r$-th region), as follows:

$$\hat{y}_r^{\mathrm{PS}} = \frac{\sum_{r \in j} w_r \hat{y}_j}{\sum_{r \in j} w_r} \tag{12}$$

This procedure ensures that when summed over the $S$ levels, the totals of $\hat{y}^{\mathrm{PS}}$ or $\hat{y}_r^{\mathrm{PS}}$ reproduce the aggregated (marginal and $k$-way) totals $N_s$ of the $K$ variables defining the population frame.

The reproduction of the population frame totals is straightforward to demonstrate. In a theoretical empirical setting, we could model OJA disaggregated by region ($r$), economic sector ($e$), occupation ($o$), and quarter ($t$) for each country individually, where the population frame, however, is provided at the level of economic sector ($e$) and quarter ($t$).

Let $\hat{y}_{\mathrm{reot}}$ be the estimated OJA at such level ($j = r \times e \times o \times t$), $N_{\bullet e \bullet t}$ be the counts of some benchmark variable for OJA in the population frame ($s = e \times t$), $\hat{y}_{\bullet e \bullet t}$ be the sum of estimated OJA at such aggregated level, and $w_{\bullet e \bullet t} = N_{\bullet e \bullet t} / \hat{y}_{\bullet e \bullet t}$ be the weights.

If we marginalize post-stratified estimates $\hat{y}^{\mathrm{PS}}_{\mathrm{reot}} = \frac{\sum_{et} w_{\bullet e \bullet t} \hat{y}_{\mathrm{reot}}}{\sum_{\mathrm{et}} w_{\bullet e \bullet t}}$ by variables in the population frame, thus by both index $r$ and $o$ ($\sum_{\mathrm{ro}} \hat{y}^{\mathrm{PS}}_{\mathrm{reot}}$), and assuming without loss of generality that the weights sum to 1 in the sample frame ($\sum_{et} w_{\bullet e \bullet t} = 1$) we obtain,

$$\begin{aligned} \sum_{ro} \hat{y}^{\mathrm{PS}}_{\mathrm{reot}} &= \sum_{ro} \left[ \frac{\sum_{et} w_{\bullet e \bullet t} \hat{y}_{\mathrm{reot}}}{\sum_{et} w_{\bullet e \bullet t}} \right] = \sum_{ro} \sum_{et} w_{\bullet e \bullet t} \hat{y}_{\mathrm{reot}} = \sum_{et} \sum_{ro} \frac{N_{\bullet e \bullet t}}{\hat{y}_{\bullet e \bullet t}} \hat{y}_{\mathrm{reot}} \\ &= \sum_{et} \frac{N_{\bullet e \bullet t}}{\hat{y}_{\bullet e \bullet t}} \sum_{ro} \hat{y}_{\mathrm{reot}} = \sum_{et} \frac{N_{\bullet e \bullet t}}{\hat{y}_{\bullet e \bullet t}} \hat{y}_{\bullet e \bullet t} = N_{\bullet e \bullet t} \end{aligned} \tag{13}$$

thus reproducing the bivariate totals of a benchmark variable for each economic sector in a given quarter within the population frame.

Another useful scenario involves post-stratifying estimates $\hat{y}_{\mathrm{reot}}$ by region and time (rt), thereby marginalizing $\hat{y}^{\mathrm{PS}}_{\mathrm{reot}}$ by economic sector and occupation (eo). In this case, since only time belongs to the population frame, this scenario simulates a population frame with only one variable (quarter $t$). Thus, the previous weights $w_{\bullet e \bullet t}$ will be marginalized also by e-levels, leading $\sum_{\mathrm{e}} w_{\bullet e \bullet t} = w_{\bullet \bullet \bullet t} = N_{\bullet \bullet \bullet t} / \hat{y}_{\bullet \bullet \bullet t}$, where $\sum_t w_{\bullet \bullet \bullet t} = 1$. Hence summing $\hat{y}^{\mathrm{PS}}_{\mathrm{reot}}$ by occupation and economic sector with new weights, gives

$$\begin{aligned} \sum_{\mathrm{eo}} \hat{y}^{\mathrm{PS}}_{\mathrm{reot}} &= \sum_{\mathrm{eo}} \sum_{t} w_{\bullet \bullet \bullet t} \hat{y}_{\mathrm{reot}} = \sum_{\mathrm{eo}} \sum_{t} w_{\bullet \bullet \bullet t} \hat{y}_{\mathrm{reot}} \\ &= \sum_{t} \frac{N_{\bullet \bullet \bullet t}}{\hat{y}_{\bullet \bullet \bullet t}} \sum_{\mathrm{eo}} \hat{y}_{\mathrm{reot}} = \sum_{t} \frac{N_{\bullet \bullet \bullet t}}{\hat{y}_{\bullet \bullet \bullet t}} \hat{y}_{\bullet \bullet \bullet t} = N_{\bullet \bullet \bullet t} \end{aligned} \tag{14}$$

reproducing the univariate totals $N_{\bullet \bullet \bullet t}$ for each quarter t across all regions.

Instead, reproducing totals $N_{r \bullet \bullet t}$ for each quarter and region is not possible, as region is not included in the population frame.

With this strategy, we can in general analyze the distributions of $\hat{y}_j^{\mathrm{PS}}$ by all $K$ variables defying the population and sample frame (as three-way or four-way tables), ensuring that their marginal and their $K$-way totals (or a subset of these $K$ variables) match the same totals of $S$ variables in the population frame.

## 4 | Data Description and Empirical Strategy

### 4.1 | OJA Data

A well-established institutional OJA repository of OJAs in Europe is the Skills-OVATE dataset (link) published by Cedefop (a decentralized EU agency that supports the European Commission in vocational education and training).

Since 2018, Cedefop started collecting data on online job advertisements in all EU member States on its web platform called the "Skills Online Vacancy Analysis Tool for Europe" (Skills-OVATE). The methodological tasks used to extract reliable information from web sources (including the selection of sources, scraping, cleaning the data, managing duplications, and text mining and classification) in standard European taxonomy have been explained elsewhere (Cedefop 2019).

Publicly available OJA are provided at an adequate level of granularity, including the occupational category (ISCO II digits), the Region (Nuts2) to which the advertisements pertain and the date of first day the OJA is posted online.

In the first empirical application we analyze Cedefop's Skills-OVATE data for all EU countries in the quarters of 2022, restricting the model fitting to Spain, Italy, Germany, and France.

The second data source refers to a well-known Italian repository that can be considered an Italian prototype of the Cedefop's Skills-OVATE source. In this end, an Italian project (2012–2013) focused on collecting online job vacancies in Italy from job portals that advertise job advertisements and include

newspaper websites, job boards, and employment agencies was carried out by Tabulex, a spin-off of the University of Bicocca-Milan, under the scientific supervision of the Interuniversity Research Centre on Public Services (CRISP), an interdisciplinary academic network of universities, led by the University of Milano-Bicocca.

Tabulex (formerly Burning Glass Europe) in 2013 created a digital solution called WOLLYBI, a Labour Market Information System with the aim of collecting, cleaning and classifying online job vacancies posted on the major Italian websites, according to methodological tasks (including selection of sources, scraping, data deduplication, text mining) of the Knowledge Discovery in Databases-KDD approach (Fayyad et al. 1996; Boselli et al. 2014, 2017; Mezzanzanica et al. 2015; Lovaglio et al. 2018; Cedefop 2019).

The portal WOLLYBI received attention from European institutions, such as Cedefop, which further initiated different pilot studies. Particularly, the project "Real-time labour market information on skill requirements: feasibility study and working prototype," funded by Cedefop, started collecting online job advertisements in all European Union member states.

Briefly, the second dataset used in the present paper refers to the Italian web platform WOLLYBI during the period 2013Q2 to 2018Q2 (21 quarters). The available data is a structured dataset where each observation represents an OJA classified according to the standard ESCO classification (ISCO-08, I, II and III digits, henceforth indicated as ESCO1, ESCO2, ESCO3, respectively), economic sector (Nace Rev.2), region (Nuts 2) and quarter.

In this end, the main advantage of using the WOLLYBI dataset lies in its provision of OJA stocks detailed across all ESCO3 × Nuts2 × Nace×Quarter levels, allowing a more nuanced analysis, whereas the public section of Skills-OVATE restricts analysis to OJA stocks disaggregated for ESCO2 × Nuts2 × Quarter (generally only five consecutive quarters).

### 4.2 | LFS Benchmark Data and Webographics

In the general framework of analysing the labour market and specifically OJA in Europe, Eurostat data is an imperative source. Official EU and extra-EU vacancy data are captured through Job Vacancy Statistics (JVS, Eurostat 2019), which are disseminated quarterly and by sector. However, JVS typically provides only the job vacancy rate for many countries (e.g., Italy, France) without specifying the actual number of positions demanded.

Given the limitations of using JVS, we compare OJA data with LFS employment data, specifically focusing on new jobs created in the past 3 months for each quarter, as suggested by recent literature (Lovaglio et al. 2018; Garasto et al. 2021; Cammeraat and Squicciarini 2021; Turrell et al. 2018).

We use two benchmark datasets from Eurostat LFS data: the first refers to "LFS recent job starters (individuals who began their employment within the 3 months preceding the interview), publicly available as quarterly series for all EU countries and sector of economic activity, from Eurostat online data (Eurostat 2024)."

For this benchmark, we downloaded data from the Eurostat website, focusing on new job starters for people aged 15–74 years in Italy from 2013-Q2 to 2018-Q2 and in the four EU countries (ES, DE, IT, FR) for the quarters of 2022. To measure the percent bias, we used only the quarterly totals of job starters as population frame, as the Nace disaggregation could not be fully utilized due to significant missing values.

As an additional benchmark dataset, we utilize the richness of LFS microdata to select a subset of new starters that aligns more closely with our objectives (LFS new hires), enabling us to focus solely on new employees and exclude unpaid workers, self-employees, or new employees in public administration.

More precisely, from quarterly LFS microdata in each country we selected employees (excluding the self-employed and unpaid family workers) who had been newly hired within the last 3 months, excluding those in "Public Administration and Defence" (ESCO1 = 00), those employed in the sector "Activities of Extraterritorial Organizations and Bodies" (Nace = U) or public administration (Nace = O). In this microdata, we exclude workers that start the work outside the country of residence and we do not consider the region of residence for these new hires, but the region of work (since OJA refer to the region where the job is vacant). By summing these counts and weighting by LFS population weights (COEFFQ), we obtained quarterly figures for LFS new hires to use as an additional benchmark.

Although the microdata could not offer a population frame, given the delay in their dissemination thus preventing a near-real-time post-stratification of OJA predictions, they are fundamental to assess the bias of OJA predictions across different statistical approaches. This is precisely how we utilize LFS microdata. Precisely, we calculated the percentage bias of models, comparing the distributions of the models' predictions with the benchmark data across the categories of main covariates: Quarters, ESCO2, ESCO1, Nuts2, and Nace. Specifically, the percentual bias was calculated as 100 × (Pred_OJA—LFS_Benchmark)/LFS_Benchmark, where LFS_Benchmark refers to both LFS benchmarks.

In modeling OJA, we also incorporate a set of covariates, particularly useful for the selection equation (instruments). These are known as webographics or attitudinal variables, as they capture differences between "online" and "offline" outcome profiles across different territorial units or economic sectors. Among these, internet penetration rates (and trends) are particularly noteworthy (Schonlau and Couper 2017; de Pedraza and Serrano 2014).

We selected yearly series of webographics variables, disaggregated at the Nuts level, for both citizens (such as internet usage, online home banking, use of social networks, etc.) and sector digitalization indicators (such as the percentage of workers employed in high-tech and knowledge-intensive occupations relative to total employment in a sector). The

variables, summarized in Table A1, are sourced from the Regional Statistics database from the "Regional Digital Economy and Society" and "Regional Science and Technology Statistics" sections (Eurostat 2023).

## 4.3 | Empirical Strategy

### 4.3.1 | OJA Manipulation

Until 2023, the Skill-OVATE repository, other than providing the number of job advertisements broken down by region (Nuts2) and occupation-ESCO2 for all EU countries, reported also the day when each OJA was downloaded by the system from online job portals (grab_date) that, considering that the data are scraped and downloaded on a daily basis, the date of scraping can be considered equivalent to the date of first day the OJA is posted online.

This is an important information since it can resolve the first main problem of OJA which is to transform OJA flows into stock. This requires assigning each Raw OJA first posting date to a specific reference quarter-year. Doing so ensures that this data can be consistently merged with other official sources that use the same reference time. Previous studies that used data from Italy, the United Kingdom, and the United States (Lovaglio et al. 2018; Lovaglio 2022; Lovaglio and Riussi 2025; Garasto et al. 2021; Cammeraat and Squicciarini 2021; Turrell et al. 2018) have shown that approximately 90% of OJAs are available for no longer than 2 months. Hence, we calculated the date corresponding to 60 days after that initially provided in the dataset (grab_date) to identify a two-month period for each OJA and we assigned each OJA to the quarter in which it remained available/open for the longest duration in the two-month period. Thus, we select OJA that refers only to quarters of 2022. The same pre-processing was done on WOLLYBI's dataset, leading OJA in the period 2013-Q2 to 2018-Q2.

### 4.3.2 | Sample Frame and Model Selection

WOLLYBI dataset was manipulated to generate possible profiles arising by cross-tabulation of Nuts2 × Nace × ESCO2 × Quarter, choosing ESCO2 as reasonably and meaningful aggregation level of analysis, whereas Skill-OVATE by Nuts2 × ESCO2 × Quarter, since the sector was not provided from publicity data. Apart from the Heckman SSM where we use the traditional logged version of the outcome (OJA counts) with bivariate Gaussian errors, for other models (ML, MGCV, GJRM) we assess the proper distribution of the outcome variable ($OJA_j$), exploring different options such as Gaussian, Lognormal, Poisson, Negative Binomial, Gamma by inspecting qq-plots.

In both (Heckman and GJRM) SSM equations and in the outcome equation of ML and MGCV we specify the covariates that define the profiles (Nuts2 and Nace only for WOLLYBI), whereas for occupation we try both ESCO1 and ESCO2. In the WOLLYBI dataset, given the large number of available quarters, the time dimension was manipulated differently in each equation (as continuous variable, as factor variable, or using the year in combination with seasonal dummies Q1-Q2-Q3-Q4), depending on the best fit provided by the different models. Obviously, the selection equation for both SSM allows the inclusion of webographics as possible instruments.

The storytelling of empirical application strongly rests on the estimation and significance of $\rho$ that will determine the most suitable methodological approach, such as whether a univariate model (e.g., ML or MGCV) or a joint model (e.g., Heckman or GJRM) is more appropriate.

For the best GJRM representation, in line with literature, we use the following strategy for choosing marginals and copulas: first we select the best univariate margins $F_1$ $F_2$ examining quantile–quantile (qq) plots on univariate outcomes distributions.

Second, we select covariates for both equations using a Gaussian copula for errors correlation. Third, different copulas C were examined, and the best copula was identified according to the lowest Bayesian Information Criterion (BIC) criterion. The models were estimated using the GJRM R package (Marra and Radice 2017). Since the estimate (and significance) of rho in the empty models would justify a bivariate classical SSM approach, apart from the GJRM package, we also run a Likelihood ratio (LR) test, testing the null that rho is zero, implemented in SAS (proc qlim) using a Newton–Raphson optimization with ridging. The estimated final (OJA) equations were evaluated using, as fit metrics, Root Mean Square Error (RMSE) and *R*-square.

## 5 | Results

### 5.1 | WOLLYBI Italian Data

In this section we start the analysis using WOLLYBI Italian data, the prototype of the Skill-Ovate European dataset. Overall, the OJA in the period from 2013-Q2 to 2018-Q2 amounted to 21,403,160. During the same period, LFS reported 15,287,000 recent job starters, whereas we selected from LFS microdata a best targeted set of 4,316,000 LFS new hires that constitutes our benchmark dataset to compare models' (MRP, MGCV, Heckman and GJRM) predictions.

These numbers indicate a significant overrepresentation of OJA compared to actual new jobs, confirming that models that do not control for sample selection and that pursue as unique aim the best fit are not appropriate since the outcome variable is strongly biased.

Since this data source is disaggregated by all relevant labor market variables, the bias assessment of models' predictions may be performed by each dimension that define the sample frame (Nace, ESCO2, Nuts2, quarter). This also allows working with an ideal sample frame where the profiles were defined by combinations of Nuts2 × ESCO2 × Nace and each profile is repeated *T* times (quarters). In this end, crossing the levels of Nuts2 (20 regions), Nace (12 aggregated levels), and different ESCO2 (43 levels), we found 10,320 profiles *J**, repeated 21 times, leading a sample frame of 216,720 rows. More than a quarter (25.4%) of such rows present a missing number of OJA, for example, an absence of online job advertisement for such combinations.

Working with a Gaussian margin for the logged version of OJA counts and a logistic margin for the selection equation (an a gaussian copula), whereas in the classic SSM we use the Probit margin for the selection equation, we found a strongly significant selection effect from both SSM ($\hat{\rho}$ =0.99***) and GJRM ($\hat{\rho}$ =0.99***), thus demonstrating that univariate approaches such as Multilevel and MGCV models lead biases and inconsistent parameters and predictions.

Particularly in the GJRM (using Nace, ESCO2, Nuts2, quarter as covariates in both equations), the hierarchical effect (profiles' dummies) is strongly significant for both equations.

Additionally, the outcome equation selects as significant covariates six webographics, such as Tech_EMPL_j and TR_Tech_EMPL_j for sectors ICT (Nace J), Manufacturing (Nace C), and Professional, Scientific and Technical Activities (Nace M).

As strongly significant instruments in the selection equation we found two nonlinear (quadratic) transformations for the percentage use of Home Bank ($z$-test = 18.7***) and for the level of digital skill of the public administration workforce (Tech_EMPL_O: $z$-test = 6.98***). These nonlinear (GAM) effects (another added value of GJRM) resulted not significant in their linear specification in the SSM-Heckman model. Thus, instruments in the Heckman model were suggested by their near quadratic specification of GJRM. The fit of the outcome equation by GJRM ($R^2$ = 0.462, RMSE = 1.446), MCGV ($R^2$ = 0.477, RMSE = 1.447) and Heckman-SSM ($R^2$ = 0.465, RMSE = 1.400) appears quite similar.

Figure 1 presents the trends over time of OJA counts, as well as two LFS benchmarks and the predictions of four models.

Firstly, the Multilevel model appears to strongly replicate OJA trajectories over time, amplifying their peaks in the predictions due to the specification of full random terms in the model, whereas MGCV, by exploiting the penalization of profiles' effects, smooths these peaks, presenting a more smoothed trajectory that aligns more closely with the benchmark (LFS new hires).

Figure 1 demonstrates that GJRM and classical SSM are better targeted to LFS new hires (whereas no one seems to "replicate" LFS recent starters over time).

Particularly, GJRM reduces the percentual bias (+43.4%) of OJA relative to the new hires benchmark (+395.9%) over the 21 quarters, performing significantly better than Heckman's SSM (+63.4%). This means that including the dummy profiles as fixed covariate in the count equation makes a good job w.r.t. to traditional SSM.

Regarding the other benchmark (LFS recent starters), the percentual bias of OJA (+40.1%) is not particularly reduced neither by SSM (−53.9%) nor by GJRM (−59.5%).

These numbers confirm that SSM in general represents a promising strategy for addressing issues of representativeness using LFS new hires as benchmark data.

Since LFS microdata provides a benchmark for all combination of Quarters × Nuts2 × Nace × ESCO2, we can control the bias of GJRM predictions for each dimension of interest. Unlikely Nuts2 and quarters where the bias is quite stable across categories, the Bias% are still problematic for some levels of occupations (ESCO1-managers +3315.5%, ESCO1-skilled agriculture/forestry/fishery +430.8%) and sectors (A-agriculture −86.3%; K-financial and insurance activities +254.4%).

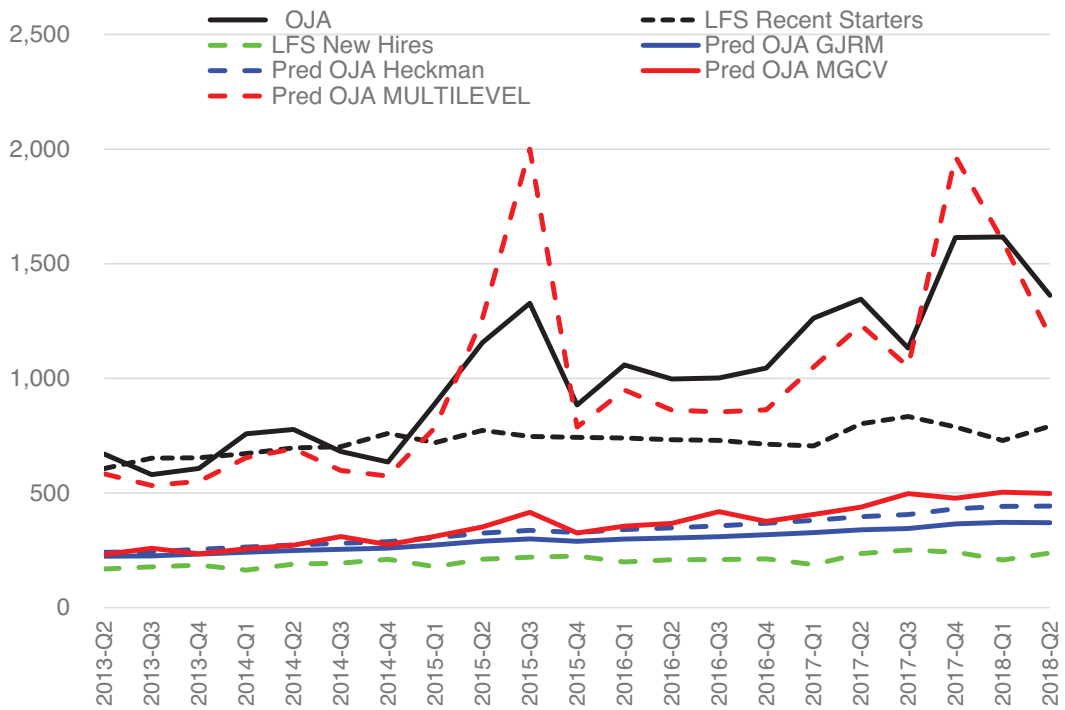


**FIGURE 1** | WOLLYBI Italian data: Evolution of OJA counts, LFS benchmarks (recent job starters and new hires) and predicted OJA by Multilevel, MGCV, Heckman-SSM, and GJRM-SSM (thousands). The Multilevel OJA Prediction at 2015-Q3 (2290) was truncated at 2000 in the figure.

### 5.2 | Cedefop Skill-Ovate Analysis of 2022

The above analysis was replicated using the most recent publicly available data (referred to 2021 and 2022) from Skills-OVATE institutional EU repository, for the four most populous EU countries. Using the *Skill-Ovate* waves that refer to five quarters of 2021–2022 (2021-Q4 to 2022-Q4) and reassigning the correct reference date to each OJA, we analyze OJA in the quarters of 2022 for all EU countries.

The levels of the sample frame were defined by combinations of Nuts2 × ESCO2 × Quarter, whereas the profiles by the combinations of Nuts2 × ESCO2 (#levels = # profiles × 4). Table 1 presents the distributions of OJA across countries (illustrating the number of Nuts2) as well as the percentage of with missing OJA.

As expected, the total OJA observed (45,681,805) were largely concentrated on a few countries such as Germany, France, Italy (52% of all OJA), Holland (and quite surprisingly Poland), whereas, quite counter-intuitively, Spain presents a very small amount of OJA.

Notice that some countries, generally those with a single Nuts2-region, present a very large percentage of selected profiles (e.g., 98.8% for LV), thus preventing a full sample selection analysis. This occurs when the obtained sample frame is not adequately disaggregated, especially in this empirical situation, where the sector was not available from OJA data.

**TABLE 1** | Number of regions (Nuts2), number of levels of the sample frame, number of profiles and profiles with non-missing OJA (selected) and OJA distributions (overall period 2022-Q1 2022-Q4) by EU country.

| | #Nuts2 | #levels of sample frame | #profiles of sample frame | %profiles selected | OJA | %OJA |
|---|---|---|---|---|---|---|
| AT | 9 | 1548 | 387 | 96.4 | 476,342 | 1.0 |
| BE | 11 | 1892 | 473 | 97.3 | 2,614,442 | 5.7 |
| BG | 6 | 1032 | 258 | 94.6 | 157,759 | 0.3 |
| CY | 1 | 172 | 43 | 95.3 | 4134 | 0.0 |
| CZ | 8 | 1376 | 344 | 96.8 | 474,660 | 1.0 |
| DE | 38 | 6536 | 1634 | 97.7 | 8,650,383 | 18.9 |
| DK | 5 | 860 | 215 | 97.2 | 159,236 | 0.3 |
| EE | 1 | 172 | 43 | 95.3 | 21,478 | 0.0 |
| EL | 13 | 2236 | 559 | 94.1 | 99,399 | 0.2 |
| ES | 18 | 3096 | 774 | 94.8 | 1,206,374 | 2.6 |
| FI | 5 | 860 | 215 | 83.3 | 57,684 | 0.1 |
| FR | 22 | 3784 | 946 | 97.7 | 11,079,959 | 24.3 |
| HR | 4 | 172 | 172 | 24.4 | 37,607 | 0.1 |
| HU | 8 | 1376 | 344 | 95.9 | 179,078 | 0.4 |
| IE | 3 | 516 | 129 | 96.9 | 401,050 | 0.9 |
| IT | 21 | 3612 | 903 | 95.7 | 4,245,438 | 9.3 |
| LT | 2 | 344 | 86 | 96.5 | 84,723 | 0.2 |
| LU | 1 | 172 | 43 | 93.0 | 2798 | 0.0 |
| LV | 1 | 172 | 43 | 100.0 | 67,066 | 0.1 |
| MT | 1 | 172 | 43 | 95.3 | 9347 | 0.0 |
| NL | 12 | 2064 | 516 | 97.7 | 3,897,393 | 8.5 |
| PL | 17 | 2924 | 731 | 97.3 | 9,779,286 | 21.4 |
| PT | 5 | 860 | 215 | 97.7 | 516,502 | 1.1 |
| RO | 8 | 1376 | 344 | 95.9 | 103,165 | 0.2 |
| SE | 8 | 1376 | 344 | 97.7 | 1,191,168 | 2.6 |
| SI | 2 | 344 | 86 | 95.3 | 23,207 | 0.1 |
| SK | 4 | 688 | 172 | 95.9 | 142,127 | 0.3 |
| Total | 231 | 39,732 | 9933 | 96.3 | 45,681,805 | 100.0 |

Before analyzing model results, Figure 2 illustrates the percentual bias for each country, measuring how the OJA distributions vary w.r.t. two benchmark LFS datasets.

Restricting the analyses to four most populous countries (DE, FR, IT, ES), from the models' analyses we excluded one ESCO2 category (ESCO63, "Subsistence Farmers, Fishers, Hunters and Gatherers") and one region (Nuts2 FRY5, Mayotte) since they were absent from both LFS microdata and OJA data.

Studying the qqplots of OJA distributions, for three countries we will work with Gaussian distributions for logged OJA, whereas Negative binomial for OJA in France. For all countries we use Gaussian copula. Table 2 illustrates the main results for each equation (covariates, penalized dummy effects that drive the time dependency, instruments of the selection equation, fit of the OJA equation), as well as the estimate of rho in a null model and the LR test.

As expected, we found a strong positive correlation coefficient between the two equations, also confirmed by the LR tests, implying a strong non-random selection, or substantial evidence against the MAR hypothesis and that the two equations cannot be modeled separately. Instead, this hypothesis cannot be refused once the covariates were considered (Final Model).

Notice the crucial role of penalized dummies (defined by Nuts1 × ESCO2 levels) that exploit the data hierarchy (time dependence) in both equations, one of the added values of the GJRM estimation approach. This is particularly useful since, in cases of absence of significant webographics in the selection equation (e.g., Italy), these "pseudo-random effects" can be used as additional instruments when they were not specified in the outcome equation. The fit is satisfactory for all countries.

Regarding the selection equation, note the strong significance of instruments (apart Italy): OJA are likely to appear online in regions where the population has an advanced "digital" skill (home banking in France) or in sectors that increased workers digital skills (in agriculture and mining, in the sector service and also in the public administration in Spain and Administrative and Support Service sector in France). In Germany, instead, regions with past increased of digital skill of workers in agriculture, mining and manufacturing are less likely to advertise online jobs in most recent years. As expected, the outcome equation strongly selects sectors with a high shares of employment in technology, especially in ICT, manufacturing and professional activities.

To conclude the analysis, the OJA predictions were compared with LFS benchmarks. These are visually represented in Figure 3.

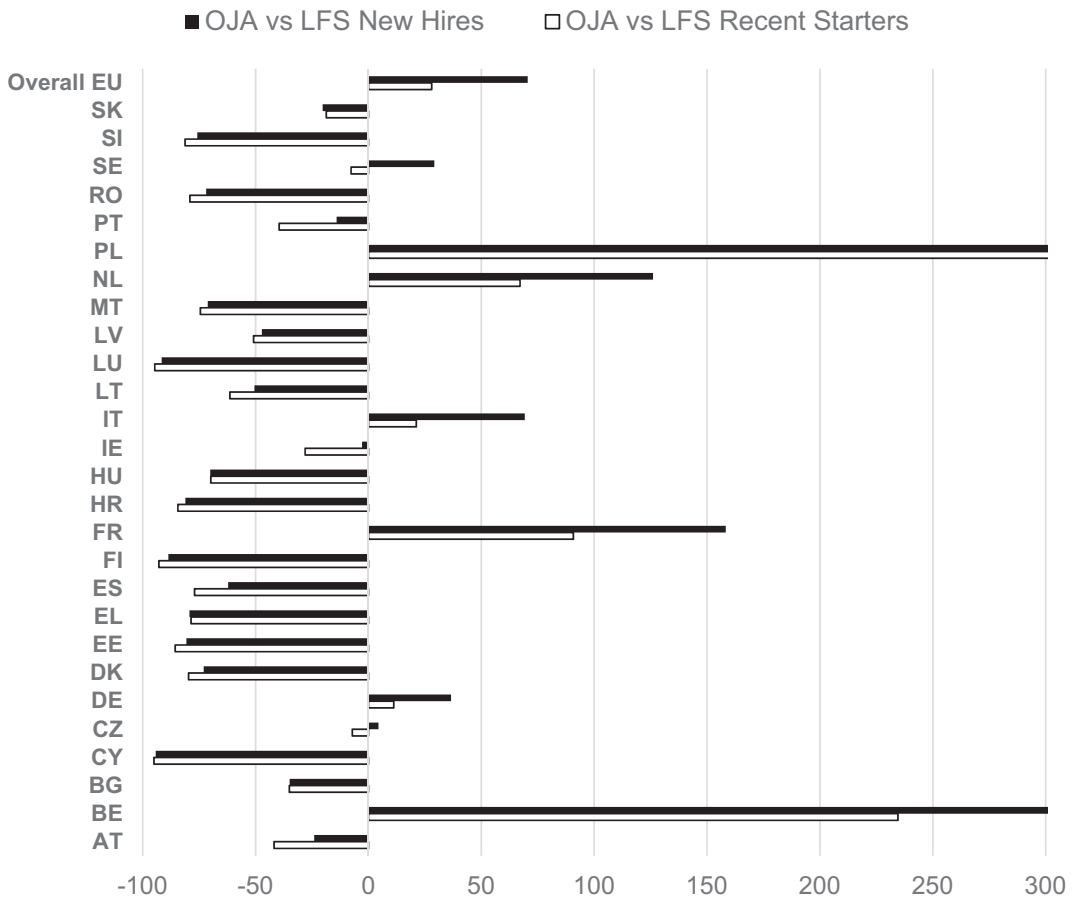


**FIGURE 2** | Bias% [(OJA-LFS)/LFS × 100] among counts of OJA, LFS new hires, and LFS Recent starters by EU countries (overall 2022 quarters). Values for Poland (Bias% = 620% w.r.t. LFS new hires and Bias% = 505% w.r.t. LFS recent starters) and Belgium (BE, Bias% = 360% w.r.t. LFS new hires) were truncated at 300%.

**TABLE 2** | Results of the GJRM models by country.

| | DE | FR[b] | IT | ES |
|---|---|---|---|---|
| Null model: Rho[a] | 0.997 | 0.999 | 0.999 | 0.999 |
| Null model: LR Test Rho = 0[a] | 237.89*** | 540.6*** | 1085.2*** | 1236.2*** |
| Final model: Rho (95% CI) | −0.060 (−0.323, 0.210) | −0.054 (−0.283, 0.146) | 0.009 (−0.147, 0.170) | 0.088 (−0.052, 0.239) |
| Selection equation | | | | |
| Outcome margin | Logit | Probit | Probit | Probit |
| Covariates[d] | D_Profile + Instruments | D_Profile + Time + Instruments[c] | D_Profile + Quarter | ESCO2 + Nuts1 + Quarter + Instruments |
| Profiles' dummies (Chi-square, sign) | Nuts1 × ESCO2 (27.71***) | Nuts1 × ESCO2 (36.53***) | Nuts1 × ESCO2 (67.82***) | |
| Instruments (coeff, signif)[e] | TR_Tech_EMPL_AB (−0.936**) | Tech_EMPL_N (0.360*) | | Tech_EMPL_AB (0.350***) |
| | TR_Tech_EMPL_C (−0.072*) | HomeBank21 (0.081***) | | EMPL_GIT (0.148***) |
| | | | | Tech_EMPL_O (0.067***) |
| OJA equation | | | | |
| Outcome margin | Log normal | Negative binomial | Log normal | Log normal |
| Covariates | ESCO2 + Nuts2 + Quarter | ESCO2 + Nuts2 + Time + D_Profile | ESCO2 + Nuts2 + Time | ESCO2 + Nuts2 + Time |
| Webographics | Tech_EMPL_J + Tech_EMPL_M | Tech_EMPL_J + Tech_EMPL_M | Tech_EMPL_J + Tech_EMPL_C | Tech_EMPL_C + Tech_EMPL_M |
| Profiles' dummies (Chi-square, sign) | — | Nuts1 × ESCO2 (401.5***) | | |
| $R^2$ (RMSE) | 0.955 (0.355) | 0.969 (0.396) | 0.947 (0.517) | 0.917 (0.626) |

[a]Empty OJA equation and using Time as unique instrument in the selection equation.
[b]Omitting Nuts2 FRY5.
[c]Time is the continuous version of the categorical variable Quarter.
[d]Sector labels for sector regarding Tech_EMPL_sector and its trend (TR_Tech_EMPL_sector) are: AB = agriculture + mining, C = manufacturing, GIT = wholesale and retail trade + accommodation and food + activities of households as employers (services), *N* = administrative and support service activities, O = public administration, J = ICT, M = professional, scientific and technical activities.
[e]Signif. codes: "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1.

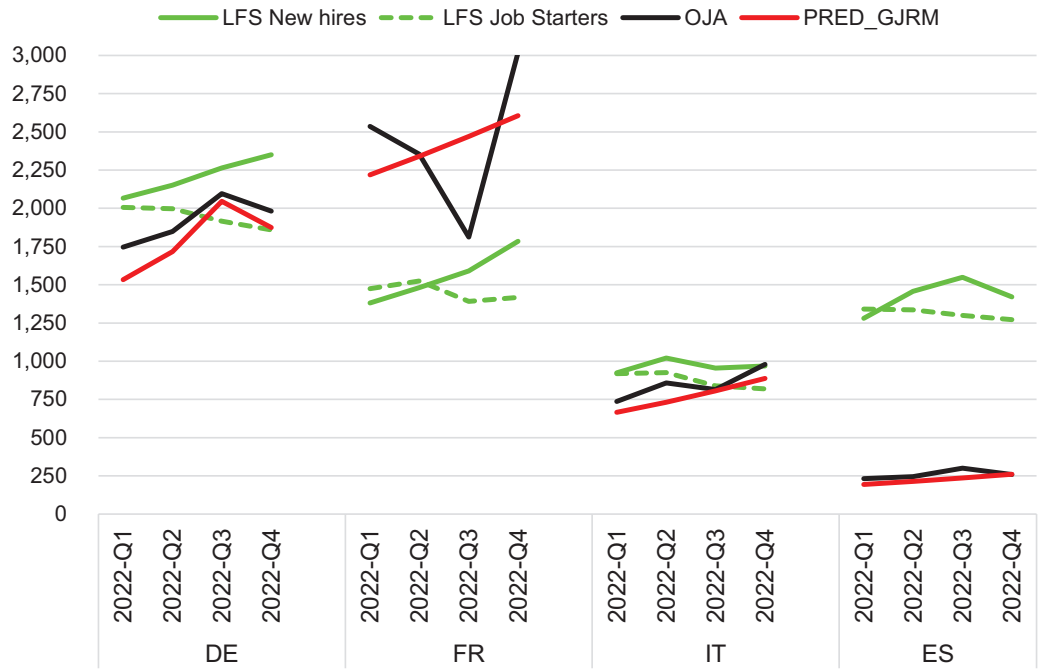


**FIGURE 3** | Cedefop EU data: Quarterly evolution in 2022 of LFS new hires (microdata), LFS recent job starters, OJA raw counts and predicted OJA by GJRM in four EU countries (thousands).

Overall, unlike OJA, the GJRM predictions are quite well targeted with benchmark data, aligning closely with LFS new hires, particularly DE and IT. However, an exception is noted for Spain, where, as previously noticed, there is a notable underrepresentation of OJA compared to both LFS benchmarks. This discrepancy may be due to the selection mechanisms of OJA on Spanish portals or unique characteristics of the Spanish labor market.

### 5.3 | Discussion

In a broader context, when the sample may not be fully representative of the population (truncated sample), we strongly recommend reorganizing the original sample data into such *sparse* cross-tabulated profiles using the auxiliary variables that we suspect drive the selection mechanism, as well as the outcome intensity.

In each specific applied context, working with sample data too much aggregated may lack profiles with missing outcome, making it impossible to use a joint sample selection model. From the other end, working with sample data too much disaggregated may present too many profiles whose sparseness may lead to insatiability of estimates of convergence problems.

The choice should converge toward an appropriate level of aggregation, mainly justified by the empirical context and by the expectation of relevance of inferences.

Regarding the use of benchmark data, three main points should be addressed.

First, it is clear that the number of new hires does not correspond to the population that the OJA sample is intended to represent. That population could be defined as the total number of OJA available on the web, or a subset thereof, such as the total vacant jobs in a given context (region, occupation, sector, or quarter).

An alternative benchmark could be the official EU data on job vacancies published by Eurostat through the Job Vacancy Statistics (JVS). This survey is representative of the stock of vacancies reported by enterprises (statistical units) at the end of each quarter, and total vacancy counts are published (although for some countries only the job vacancy rate is available).

However, JVS data lack the level of granularity required for use as a benchmark in our context. They are collected through quarterly surveys, typically disaggregated only at the sectoral level (with data availability varying across EU countries), and provide no occupational detail and no regional information at the NUTS2 level. As a result, JVS data allow only a limited assessment of underlying labor market conditions and are not suitable for near real-time monitoring.

We therefore treat the adopted benchmark—the quarterly stock of new jobs created over the last 3 months, a well-defined and readily available measure from the LFS—as a lower bound for the population that OJA data are intended to represent. This is because not all posted vacancies necessarily result in hires, and thus the number of new hires can only approximate a subset of the total vacancy stock captured by OJA. This choice agrees with previous studies that assess representativeness of OJA (Cammeraat and Squicciarini 2021; Garasto et al. 2021). In this perspective, if the benchmark were defined instead as the stock of jobs actually created in the past 3 months—as in our case (a still narrower subset of OJA)—an adequate and proper way to evaluate model performance would be to calculate the number of new hires implied by the predicted vacancy counts under a given job filling rate.

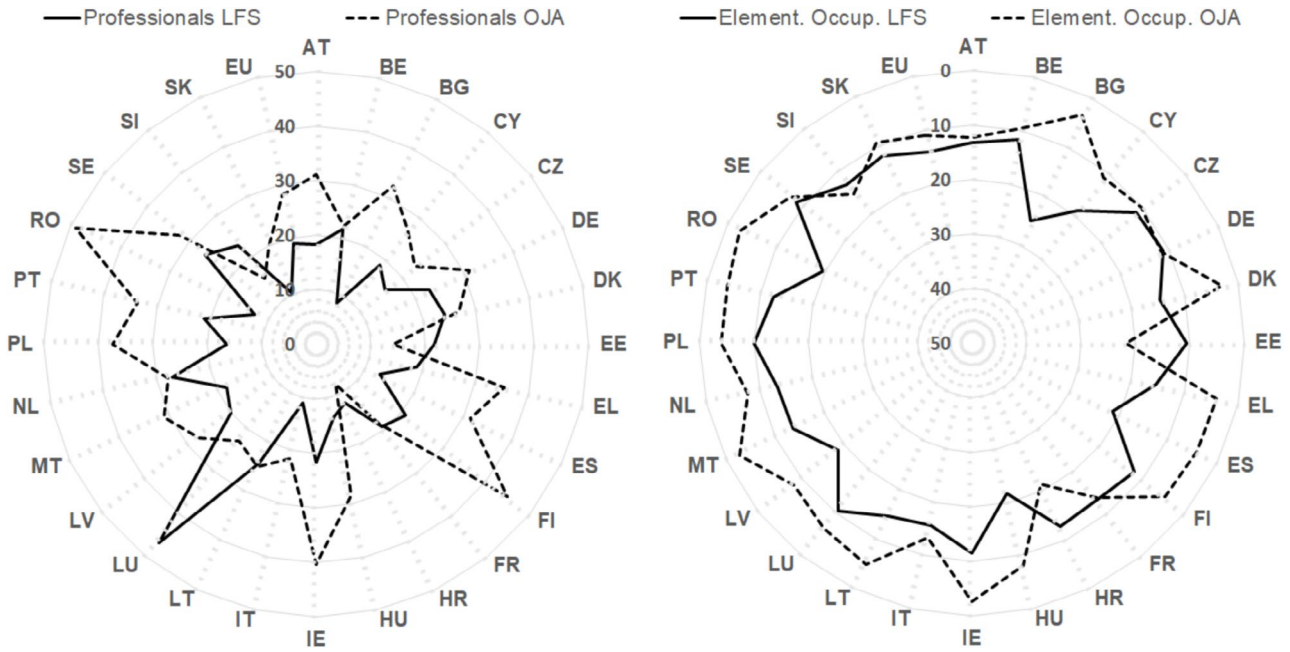


**FIGURE 4** | Shares of professionals (left) and elementary occupations (right) based on OJA counts and LFS new hires in 2022-Q4, by country.

However, while the fill rate is well defined in traditional labor market statistics, it is not usually computed for OJA data. According to the US Bureau of Labor Statistics, the fill rate is calculated as the ratio of hires in a month or quarter to job openings at the end of the previous month or quarter. To date, there is no widely adopted or standardized estimate of the fill rate specifically for OJA data, and no authoritative sources currently report such a measure.

This limitation supports the use of LFS new hires as a reliable, though imperfect, benchmark. Accordingly, the bias measure should be interpreted with caution. In particular, %Bias should not be viewed as a strict measure of deviation from the benchmark, but rather as a comparative metric to evaluate differences between predicted OJA and this lower bound.

Second, since the main purpose of benchmark data is to recalibrate model predictions of online data through post-stratification, their key requirement is availability in near real time. As an example, it is still quite clear that since LFS microdata are released with a delay of nearly 2 years, they are unsuitable for near-real-time OJA. On the other end, quarterly LFS data on recent job starters remains a uniquely valuable resource for adjusting OJA data in near-real time.

Third, benchmark data should be disseminated at an acceptable disaggregation level or by relevant auxiliary variables in each empirical context.

Continuing with the previous example, LFS recent job starters are disaggregated only by Quarter and Nace level (also containing numerous missing values, varying by EU country), missing important information mediated by occupation type (ESCO).

In this end, exploiting the fact that our data refer to a window for which LFS microdata are available, we assess potential misalignment between the observed OJA and LFS New hires by occupations, a dimension not available in LFS Recent Starters. Among various comparisons we provide how shares of most extreme occupations vary among OJA and LFS new hires in all EU countries (Figure 4).

Moreover, the structure and the information content of OJA is such that classification into occupation is more precise than classification into sectors of economic activity. In fact, often OJA report the sector of the enterprise that posts the advertisement, rather than the sectors of the required job. This point opens the question of possible measurement error bias that should be explored in combination with non-response bias (Zhang 2005).

Overall, such considerations claim that to accurately adjust OJA stocks through post-stratification, it is essential to have near-real time benchmark data also at ESCO1-level disaggregation. Otherwise, this limitation prevents precise real-time adjustments of OJA for labor demand analyses.

On the other end, the portal Cedefop Skill-Ovate data, since the beginning of 2023 (Q1) no longer provides the grab date for OJA data, which makes it impossible to separate OJA by quarters and replicate our analysis for more recent periods.

Accordingly, the above discussion highlights several limitations of the paper.

## 6 | Conclusion

The present paper has illustrated possible strategies to deal with the problem of representativeness of OJA and to align these data with more common official statistics used in the labor market.

Model-based inference and post-stratification are standard techniques for addressing unequal probabilities of selection and non-response, provided the non-response mechanism is known

or partially accounted for by covariates. However, apart from a few exceptions, these approaches including one of the more popular frameworks, such as Multilevel regression and post-stratification, largely work on the assumption of Missing at Random mechanism.

Even if our proposed strategy is quite different from classical MRP, it is inspired by the idea to estimate the bivariate SSM in the sample frame dataset. In this end, the core of the methodological proposal is the creation of the sparse sample frame, derived from the cross-tabulation of the levels of auxiliary variables and obtained exactly in the same manner as the population frame.

The presented approach can be seen as a doubly robust methodology dealing with self-selection as it exploits both a model-based approach (typically yielding smaller non-response bias for estimator of target variable totals, also in the presence of MNAR) and a post-stratification strategy (leading always to variance reduction in case of a high association between auxiliary variables of the population frame and the outcome).

Indeed, literature on non-random selectivity and post-stratification suggests that for the validity of MAR, the focus should be on the high predictive power of covariates selected for the outcome equation.

In particular, strong covariates in the outcome equation lead always to variance reduction and sometimes to non-response bias reduction, whereas the propensity model can be helpful in terms of bias reduction *only* if there is a high association between such covariates and the outcome. In the same perspective, it is not useful to incorporate auxiliary variables in the population frame that are not predictive of the outcome variable, even though they may have a high association with non-response (Little and Vartivarian 2005; Buelens et al. 2018; Matei 2018; Zhang et al. 2013).

In other words, the explanatory power of auxiliary variables for the propensity equation is not helpful without the prediction power of such variables also for the outcome variable.

However, we doubt that strategies that do not focus on controlling sample selection are pursuing a correct strategy, such as merely enhancing the predictive power of a fitted model, given that the observed outcomes stem from a truncated and non-representative sample. Indeed, our approach seems more realistic, since it does not hypothesize an a priori mechanism for missingness, but establishes it based on empirical estimations (as we found in the EU application, where Italy demonstrates a MAR mechanism, once the observed covariates were taken into account, unlike the other three countries).

Moreover, a new weighting strategy, ensuring that finer level distributions of predicted outcomes, once aggregated, are constrained to match the marginal or bivariate distributions of auxiliary variables in the population frame, is proposed.

The approach presented is straightforward and, aside from an initial phase of data manipulation, can be easily implemented using standard statistical software.

Beyond the labor market, this methodology could be extended to other application areas, such as political polls or other non-representative surveys. For example, it could be used to estimate voting patterns or public opinion on ethical issues.

As far as empirical results are concerned, the paper disseminates many insights from online job adverts within a large European international project involving four populous EU countries. We use millions of job adverts to provide granular estimates of OJA distributions broken down by location, occupation, and sector also offering evidence on the relationship between OJA and official statistics.

To conclude, harmonizing online web data with (the aid of) official statistics to provide measurements at relatively lower cost and with higher frequency appears to be the main methodological challenge for future labor market analyses.


### Funding

The authors have nothing to report.

### Acknowledgements

Open access publishing facilitated by Universita degli Studi di Milano-Bicocca, as part of the Wiley - CRUI-CARE agreement.


### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

### References


Askitas, N., and K. F. Zimmermann. 2009. "Google Econometrics and Unemployment Forecasting." *Applied Economics Quarterly* 55, no. 2: 107–120.

Askitas, N., and K. F. Zimmermann. 2015. "The Internet as a Data Source for Advancement in Social Sciences." *International Journal of Manpower* 36, no. 1: 2–12.

Beręsewicz, M., R. Lehtonen, F. Reis, L. Di Consiglio, and M. Karlberg. 2018. *An Overview of Methods for Treating Selectivity in Big Data Sources. Statistical Working Papers*. Eurostat.

Boehmke, F. J., D. S. Morey, and M. Shannon. 2006. "Selection Bias and Continuous-Time Duration Models: Consequences and a Proposed Solution." *American Journal of Political Science* 50: 192–207. https://doi.org/10.1111/j.1540-5907.2006.00178.x.

Bokányi, E., Z. Lábszki, and G. Vattay. 2017. "Prediction of Employment and Unemployment Rates From Twitter Daily Rhythms in the US." *EPJ Data Science* 6, no. 1: 14.

Boselli, R., M. Cesarini, F. Mercorio, and M. Mezzanzanica. 2014. "Planning Meets Data Cleansing." In *The 24th International Conference on Automated Planning and Scheduling (ICAPS)*, 439–443. PKP Publishing Services Network.

Boselli, R., M. Cesarini, F. Mercorio, and M. Mezzanzanica. 2017. "Using Machine Learning for Labour Market Intelligence." In *Machine Learning and Knowledge Discovery in Databases*, edited by Y. Altun. Springer.

Box-Steffensmeier, J. M., and B. D. Jones. 2004. *Event History Modeling: A Guide for Social Sciences*. Cambridge University Press.

Boyes, W. J., D. L. Hoffman, and S. A. Low. 1989. "An Econometric Analysis of the Bank Credit Scoring Problem." *Journal of Econometrics* 40: 3–14.

Broniecki, P., L. Leemann, and R. Wüest. 2021. "Improved Multilevel Regression With Poststratification Through Machine Learning." *Journal of Politics* 84, no. 2: 597–601.

Buelens, B., J. Burger, and J. A. van den Brakel. 2018. "Comparing Inference Methods for Non-Probability Samples." *International Statistical Review* 86, no. 2: 322–343.

Buttice, M. K., and B. Highton. 2013. "How Does Multilevel Regression and Poststratification Perform With Conventional National Surveys?" *Political Analysis* 21, no. 4: 449–467.

Cammeraat, E., and M. Squicciarini. 2021. "Burning Glass Technologies' Data Use in Policy-Relevant Analysis: An Occupation-Level Assessment." In *OECD Science, Technology and Industry Working Papers, No. 2021/05*. OECD Publishing. https://doi.org/10.1787/cd75c3e7-en.

Caperna, G., M. Colagrossi, A. Geraci, and G. Mazzarella. 2020. "Googling Unemployment During the Pandemic: Inference and Nowcast Using Search Data." In *JRC Working Papers in Economics and Finance 2020-04*. Joint Research Centre, European Commission Publisher.

Cedefop. 2019. "Project "Real-Time Labour Market Information on Skill Requirements: Feasibility Study and Working Prototype"." https://www.cedefop.europa.eu/en/about-cedefop/public-procurement/real-time-labour-market-information-skill-requirements-feasibility.

Cedefop. 2023. *Skills in Transition: The Way to 2035*. Publications Office of the European Union. https://doi.org/10.2801/438491.

Cedefop. 2024. "Untangling Labour Shortages in Europe: Unmet Skill Demand or Bad Jobs?" https://doi.org/10.2801/023297.

Chib, S., E. Greenberg, and I. Jeliazkov. 2009. "Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection." *Journal of Computational and Graphical Statistics* 18: 321–348.

Choi, H., and H. Varian. 2012. "Predicting the Present With Google Trends." *Economic Record* 88, no. 1: 2–9.

Couper, M. P. 2013. "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys." *Survey Research Methods* 7, no. 3: 145–156.

de Pedraza, P., and F. Serrano. 2014. "Mind the Gap in the Digital Data Tsunami." In *Working Paper 154, Amsterdam Institute for Advanced Labour Studies*. Amsterdam University Press.

de Pedraza, P., S. Visintin, K. Tijdens, and G. Kismihók. 2019. "Survey vs Scraped Data: Comparing Time Series Properties of Web and Survey Vacancy Data." *IZA Journal of Labor Economics* 8, no. 1: 1–23. https://doi.org/10.2478/izajole-2019-0004.

Deming, D., and L. B. Kahn. 2018. "Skill Requirements Across Firms and Labour Markets: Evidence From Job Postings for Professionals." *Journal of Labor Economics* 36, no. S1: S337–S369. https://doi.org/10.1086/694106.

Diamond, P., and A. Şahin. 2014. "Shifts in the Beveridge Curve." In *Federal Reserve Bank of New York Staff Reports*, vol. 687. Federal Reserve Bank of New York.

Elliott, M. R., and R. Valliant. 2017. "Inference for Nonprobability Samples." *Statistical Science* 32, no. 2: 249–264.

Eurostat. 2019. "Job Vacancy Statistics." https://ec.europa.eu/eurostat/cache/metadata/en/jvs_esms.htm.

Eurostat. 2021. "LFS Main Indicators." https://ec.europa.eu/eurostat/cache/metadata/en/lfsi_esms.htm.

Eurostat. 2023. "Labour Market Statistics at Regional Level." https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Labour_market_statistics_at_regional_level#Employment.

Eurostat. 2024. "Recent Job Starters by Sex and Age - Quarterly Data." https://ec.europa.eu/eurostat/databrowser/view/lfsi_sta_q__custom_12087241/default/table?lang=en.

Fan, J., F. Han, and H. Liu. 2014. "Challenges of Big Data Analysis." *National Science Review* 1: 293–314.

Fan, J., and L. Liao. 2012. "Endogeneity in Ultrahigh Dimension." *Annals of Statistics* 42, no. 3: 872–917.

Faryna, O., T. Pham, O. Talavera, and A. Tsapin. 2022. "Wage and Unemployment: Evidence From Online Job Vacancy Data." *Journal of Comparative Economics* 50, no. 1: 52–70.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. "The KDD Process for Extracting Useful Knowledge From Volumes of Data." *Communications of the ACM* 39, no. 11: 27–34.

Fondeur, Y., and F. Karamé. 2013. "Can Google Data Help Predict French Youth Unemployment?" *Economic Modelling* 30, no. 1: 117–125. https://doi.org/10.1016/j.econmod.2012.07.017.

Garasto, S., J. Djumalieva, K. Kanders, R. Wilcock, and C. Sleeman. 2021. "Developing Experimental Estimates of Regional Skill Demand (ESCoE DP 2021-02)." ESCoE Discussion Paper 2021-02.

Gelman, A. 2007. "Struggles With Survey Weighting and Regression Modeling." *Statistical Science* 22, no. 2: 153–164.

Gelman, A. 2014. "How Bayesian Analysis Cracked the Red-State, Blue-State Problem." *Statistical Science* 29, no. 1: 26–35. http://www.jstor.org/stable/43288447.

Gelman, A., J. Lax, J. Phillips, J. Gabry, and R. Trangucci. 2016. "Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion. Unpublished Manuscript." http://www.columbia.edu/~jhp2121/workingpapers/MRT.pdf.

Gelman, A., D. Lee, and Y. Ghitza. 2010. "Public Opinion on Health Care Reform." *Forum* 8: 1–14.

Gelman, A., and T. C. Little. 1997. "Poststratification Into Many Categories Using Hierarchical Logistic Regression." *Survey Methodology* 23, no. 2: 127–135.

Gelman, A., D. Park, B. Shor, and J. Cortina. 2009. *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*. Princeton University Press.

Gelman, A., B. Shor, J. Bafumi, and D. Park. 2007. "Rich State, Poor State, Red State, Blue State: What's the Matter With Connecticut?" *Quarterly Journal of Political Science* 2: 345–367.

Ghitza, Y., and A. Gelman. 2013. "Deep Interactions With MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science* 57, no. 3: 762–776. https://doi.org/10.1111/ajps.12004.

Greene, W. 2001. "FIML Estimation of Sample Selection Models for Count Data." In *Economic Theory, Dynamics and Markets. Research Monographs in Japan-U.S. Business and Economics*, edited by T. Negishi, R. V. Ramachandran, and K. Mino, vol. 5. Springer.

Greene, W. H. 2012. *Econometric Analysis*. Prentice Hall.

Heckman, J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5: 475–492.

Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47, no. 1: 153–161. https://doi.org/10.2307/1912352.

Hershbein, B., and L. B. Kahn. 2018. "Do Recessions Accelerate Routine-Biased Technological Change? Evidence From Vacancy

Postings." *American Economic Review* 108, no. 7: 1737–1772. https://doi.org/10.1257/aer.20161570.

Hobijn, B., and A. Şahin. 2013. "Beveridge Curve Shifts Across Countries Since the Great Recession." *IMF Economic Review* 61, no. 4: 566–600.

Japec, L., F. Kreuter, M. Berg, et al. 2015. "American Association for Public Opinion Research (AAPOR) Report on Big Data." http://www.aapor.org/Education-Resources/Reports/Big-Data.aspx.

Kenkel, D., and J. Terza. 2001. "The Effect of Physician Advice on Alcohol Consumption: Count Regression With an Endogenous Treatment Effect." *Journal of Applied Econometrics* 16: 165–184.

Kreuter, F., and R. D. Peng. 2014. "Extracting Information From Big Data: Issues of Measurement, Inference and Linkage." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by J. Lane, V. Stodden, S. Benden, and H. Nissenbaum, 257–275. Cambridge University Press.

Kruskal, W., and F. Mosteller. 1979. "Representative Sampling, III: The Currents Statistical Literature." *International Statistical Review* 47, no. 3: 245–265.

Lax, J. R., and J. H. Phillips. 2009. "How Should We Estimate Public Opinion in the States?" *American Journal of Political Science* 53, no. 1: 107–121.

Lax, J. R., and J. H. Phillips. 2013. *How Should we Estimate Sub-National Opinion With MRP* Unpublished MS. Columbia University Press.

Lee, L. F. 1983. "Generalized Econometric Models With Selectivity." *Econometrica* 51: 507–512.

Little, R. J. A. 1993. "Pattern-Mixture Models for Multivariate Incomplete Data." *Journal of the American Statistical Association* 88: 125–134.

Little, R. J. A. 2007. "Comment: Struggles With Survey Weighting and Regression Modeling." *Statistical Science* 22, no. 2: 171–174.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis With Missing Data*. 2nd ed. Wiley.

Little, R. J. A., and S. Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31: 161–168.

Lopez-Martin, J., J. H. Phillips, and A. Gelman. 2022. "Multilevel Regression and Poststratification Case Studies." https://bookdown.org/jl5522/MRP-case-studies/.

Lovaglio, P. G. 2022. "Do Job Vacancies Variations Anticipate Employment Variations by Sector? Some Preliminary Evidence From Italy." *Labour* 36: 71–93. https://doi.org/10.1111/labr.12213.

Lovaglio, P. G., M. Cesarini, F. Mercorio, and M. Mezzanzanica. 2018. "Skills in Demand for ICT and Statistical Occupations: Evidence From Web Vacancies." *Statistical Analysis and Data Mining: Theoretical and Applied* 2, no. 11: 78–91.

Lovaglio, P. G., M. Mezzanzanica, and E. Colombo. 2020. "Comparing Time Series Characteristics of Official and Web Job Vacancy Data." *Quality & Quantity* 54, no. 1: 85–98. https://doi.org/10.1007/s11135-019-00940-3.

Lovaglio, P. G., and M. Riussi. 2025. "How Is Labour Demand Changing Across European Regions in the Post-COVID-19 Era." *International Regional Science Review* 48, no. 5-6: 524–575.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Economics*. Cambridge University Press.

Marinescu, I., and R. Rathelot. 2018. "Mismatch Unemployment and the Geography of Job Search." *American Economic Journal: Macroeconomics* 10, no. 3: 42–70.

Marinescu, I., and R. Wolthoff. 2020. "Opening the Black Box of the Matching Function: The Power of Words." *Journal of Labor Economics* 38, no. 2: 535–568.

Marra, G., and R. Radice. 2011. "Estimation of a Semiparametric Recursive Bivariate Probit Model in the Presence of Endogeneity." *Canadian Journal of Statistics* 39: 259–279.

Marra, G., and R. Radice. 2013. "Estimation of a Regression Spline Sample Selection Model." *Computational Statistics & Data Analysis* 61: 158–173.

Matei, A. 2018. "On Some Reweighting Schemes for Nonignorable Unit Nonresponse." *Survey Statistician* 77: 21–33.

Mezzanzanica, M., R. Boselli, M. Cesarini, and F. Mercorio. 2015. "A Model-Based Evaluation of Data Quality Activities in KDD." *Information Processing and Management* 51, no. 2: 144–166.

Miranda, A. 2004. "FIML Estimation of an Endogenous Switching Model for Count Data." *Stata Journal: Promoting Communications on Statistics and Stata* 4: 40–49.

Miranda, A., and S. Rabe-Hesketh. 2006. "Maximum Likelihood Estimation of Endogenous Switching and Sample Selection Models for Binary, Ordinal, and Count Variables." *Stata Journal: Promoting Communications on Statistics and Stata* 6, no. 3: 285–308.

National Science Foundation (NSF). 2018. "Collaborative Research: Multilevel Regression and Poststratification: A Unified Framework for Survey Weighted Inference." https://www.nsf.gov/awardsearch/showAward?AWD_ID=1760133.

Norrander, B. 2007. "Choosing Among Indicators of State Public Opinion." *State Politics and Policy Quarterly* 7, no. 2: 152–159.

Park, D. K., A. Gelman, and J. Bafumi. 2006. "State Level Opinions From National Surveys: Poststratification Using Multilevel Logistic Regression." In *Public Opinion in State Politics*, edited by J. E. Cohen, 35–50. Stanford University Press.

Pfeffermann, D. 2007. "Comment: Struggles With Survey Weighting and Regression Modeling." *Statistical Science* 22, no. 2: 179–183.

Prieger, J. E. 2002. "A Flexible Parametric Selection Model for Non-Normal Data With Application to Health Care Usage." *Journal of Applied Econometrics* 17, no. 4: 367–392.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2004. "Generalized Multilevel Structural Equation Modelling." *Psychometrika* 69, no. 2: 167–190.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2005. "Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models With Nested Random Effects." *Journal of Econometrics* 128: 301–323. https://doi.org/10.1016/j.jeconom.2004.08.017.

Robertson, A. 1955. "Prediction Equations in Quantitative Genetics." *Biometrics* 11: 95–98.

Ruppert, D., M. P. Wand, and R. J. Carroll. 2003. *Semiparametric Regression*. Cambridge University Press.

Şahin, A., J. Song, G. Topa, and G. L. Violante. 2014. "Mismatch Unemployment." *American Economic Review* 104, no. 11: 3529–3564.

Schmidt, T., and S. Vosen. 2013. "Using Internet Data to Account for Special Events in Economic Forecasting." *Ruhr Economic Paper* 382: 1–17.

Schonlau, M., and M. P. Couper. 2017. "Options for Conducting Web Surveys." *Statistical Science* 32, no. 2: 279–292. https://doi.org/10.1214/16-STS597.

Smith, M. D. 2003. "Modelling Sample Selection Using Archimedean Copulas." *Econometrics Journal* 6, no. 1: 99–123.

Štefánik, M., Š. Lyócsa, and M. Bilka. 2022. "Using Online Job Postings to Predict Key Labour Market Indicators." *Social Science Computer Review* 41, no. 5: 1630–1649.

Tam, S. M., and F. Clarke. 2015. "Big Data, Official Statistics, and Some Initiatives by the Australian Bureau of Statistics." *International Statistical Review* 83: 436–448.

Terza, J. V. 1998. "Estimating Count Data Models With Endogenous Switching: Sample Selection and Endogenous Treatment Effects." *Journal of Econometrics* 84: 129–154.

Trivedi, P. K., and D. M. Zimmer. 2007. "Copula Modeling: An Introduction for Practitioners." *Foundations and Trends in Econometrics* 1, no. 1: 1–111.

Turrell, A., B. J. Speigner, J. Djumalieva, D. Copple, and J. Thurgood. 2019. "Transforming Naturally Occurring Text Data Into Economic Statistics: The Case of Online Job Vacancy Postings." In *Big Data for 21st Century Economic Statistics*, edited by K. G. Abraham, R. S. Jarmin, B. Moyer, and M. D. Shapiro. University of Chicago Press.

Turrell, A., J. Thurgood, D. Copple, J. Djumalieva, and B. Speigner. 2018. *Staff Working Paper No. 742: Using Online Job Vacancies to Understand the UK Labour Market From the Bottom-Up*. Bank of England. www.bankofengland.co.uk/working-paper/staff-working-papers.

Valliant, R. 2019. "Comparing Alternatives for Estimation From Nonprobability Samples." *Journal of Survey Statistics and Methodology* 8: 231–263. https://doi.org/10.1093/jssam/smz003.

Valliant, R., A. H. Dorfman, and R. M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley & Sons.

Wang, W., D. Rothschild, S. Goel, and A. Gelman. 2014. "Forecasting Elections With Non-Representative Polls." *International Journal of Forecasting* 31, no. 3: 980–991.

Warshaw, C., and J. Rodden. 2012. "How Should We Measure District-Level Public Opinion on Individual Issues?" *Journal of Politics* 74, no. 1: 203–219.

Winkelmann, R. 1998. "Count Data Models With Selectivity." *Econometric Reviews* 17: 339–359.

Winkelmann, R. 2011. "Copula Bivariate Probit Models: With an Application to Medical Expenditures." *Health Economics* 21: 1444–1455.

Wood, S. N. 2004. "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models." *Journal of the American Statistical Association* 99, no. 467: 673–686.

Wood, S. N. 2011. "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." *Journal of the Royal Statistical Society: Series B* 73, no. 1: 3–36.

Wu, M. C., and R. J. Carroll. 1988. "Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process." *Biometrics* 44: 175–188.

Wyszynski, K., and G. Marra. 2018. "Sample Selection Models for Count Data in R." *Computational Statistics* 33: 1385–1412. https://doi.org/10.1007/s00180-017-0762-y.

Zhang, L. C., I. Thomsen, and Ø. Kleven. 2013. "On the Use of Auxiliary and Paradata for Dealing With Non-Sampling Errors in Household Surveys." *International Statistical Review* 81, no. 2: 270–288.

Zhang, L.-C. 2005. "On the Bias in Gross Labour Flow Estimates due to Nonresponse and Misclassification." *Journal of Official Statistics* 21: 591–604.

Zhelonkin, M., M. G. Genton, and E. Ronchetti. 2016. "Robust Inference in Sample Selection Models." *Journal of the Royal Statistical Society: Series B* 78: 805–827. https://doi.org/10.1111/rssb.12136.

Zilian, L. S., S. S. Zilian, and G. Jäger. 2021. "Labour Market Polarisation Revisited: Evidence From Austrian Vacancy Data." *Journal for Labour Market Research* 55: 7. https://doi.org/10.1186/s12651-021-00290-4.

## Appendix A

**TABLE A1** | Webographics from Regional Statistics database (link).

| GoodsServices_WEB | % of individuals who online purchase in the last 3 months (378 Nuts2 annual series 2006–2023) |
|---|---|
| REGULAR_WEB | % of individuals regularly using the internet (214 Nuts2, annual series 2012–23) |
| HOME_WEB | % Households with internet access at home (378 Nuts2, annual series 2006–23) |
| HOME_BANK | % of individuals who used the internet for home banking (378 Nuts2 annual series 2006–2023) |
| SOCIAL_WEB | Percentage of Individuals who used the internet for social networks, such as creating user profile, posting, sharing etc. (378 Nuts2 annual series 2006–23) |
| Interact_WEB | % of individuals who used the internet for interaction with public authorities (359 Nuts2, annual series 2008–21) |
| NEVER_WEB | % of who have never used a computer (350 Nuts2 annual series 2006–2017) |
| MOBILE_WEB | %of individuals who accessed the internet away from home or work (357 Nuts2, annual series 2012–19) |
| Tech_EMPL_j | %of employment in technology and knowledge-intensive occupations on total employment for sector j in 2021 (505 Nuts2, annual series 2013–22) |
| TR_Tech_EMPL_j | Quarterly trend of Tech_Empl_j over years 2018–2021 |

**GJRM Models Results**

**Germany**

Residual rho (95% Conf.Int.) = −0.0602(−0.323, 0.210).

| Selection equation (Logit Margin) | | | | | |
|---|---|---|---|---|---|
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | **Sign** |
| Intercept | 6.323 | 0.330 | 19.18 | <2e−16 | *** |
| TR_Tech_EMPL_AB | −0.936 | 0.330 | −2.83 | 0.0046 | ** |
| TR_Tech_EMPL_C | −0.073 | 0.030 | −2.40 | 0.0163 | * |
| Profiles'dummies (Nuts1 × ESCO2) Chi-sq. | | | 106.10 | <2e−16 | *** |
| **OJA equation (logOJA-Gaussian Margin)** | | | | | |
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | **Sign** |
| Intercept | 7.078 | 0.040 | 176.59 | <2e−16 | *** |
| ESCO2_12 | 0.939 | 0.040 | 23.57 | <2e−16 | *** |
| ESCO2_13 | −0.334 | 0.040 | −8.40 | <2e−16 | *** |
| ESCO2_14 | −0.090 | 0.040 | −2.25 | 0.0242 | * |
| ESCO2_21 | 1.496 | 0.040 | 37.57 | <2e−16 | *** |
| ESCO2_22 | 0.401 | 0.040 | 10.08 | <2e−16 | *** |
| ESCO2_23 | 0.277 | 0.040 | 6.95 | 0.0000 | *** |
| ESCO2_24 | 1.650 | 0.040 | 41.42 | <2e−16 | *** |
| ESCO2_25 | 1.923 | 0.040 | 48.29 | <2e−16 | *** |
| ESCO2_26 | 0.127 | 0.040 | 3.18 | 0.0015 | ** |
| ESCO2_31 | 1.198 | 0.040 | 30.08 | <2e−16 | *** |
| ESCO2_32 | 0.387 | 0.040 | 9.71 | <2e−16 | *** |
| ESCO2_33 | 1.754 | 0.040 | 44.05 | <2e−16 | *** |
| ESCO2_34 | 0.131 | 0.040 | 3.29 | 0.0010 | ** |
| ESCO2_35 | −0.390 | 0.040 | −9.80 | <2e−16 | *** |
| ESCO2_41 | 0.003 | 0.040 | 0.08 | 0.9402 | |
| ESCO2_42 | 0.724 | 0.040 | 18.19 | <2e−16 | *** |
| ESCO2_43 | 0.457 | 0.040 | 11.48 | <2e−16 | *** |
| ESCO2_44 | 0.400 | 0.040 | 10.04 | <2e−16 | *** |
| ESCO2_51 | 0.604 | 0.040 | 15.17 | <2e−16 | *** |
| ESCO2_52 | 1.238 | 0.040 | 31.09 | <2e−16 | *** |
| ESCO2_53 | 0.728 | 0.040 | 18.28 | <2e−16 | *** |
| ESCO2_54 | −1.624 | 0.040 | −40.78 | <2e−16 | *** |
| ESCO2_61 | −2.371 | 0.040 | −59.55 | <2e−16 | *** |
| ESCO2_62 | −4.423 | 0.040 | −110.29 | <2e−16 | *** |
| ESCO2_71 | 0.376 | 0.040 | 9.44 | <2e−16 | *** |
| ESCO2_72 | 0.980 | 0.040 | 24.61 | <2e−16 | *** |
| ESCO2_73 | −1.432 | 0.040 | −35.96 | <2e−16 | *** |
| ESCO2_74 | 0.613 | 0.040 | 15.39 | <2e−16 | *** |
| ESCO2_75 | −0.458 | 0.040 | −11.50 | <2e−16 | *** |
| ESCO2_81 | 0.559 | 0.040 | 14.05 | <2e−16 | *** |
| ESCO2_82 | 0.465 | 0.040 | 11.69 | <2e−16 | *** |

| | OJA equation (logOJA-Gaussian Margin) | | | | |
|---|---|---|---|---|---|
| | Estimate | Std.Error | *z* value | Pr(>\|z\|) | Sign |
| ESCO2_83 | 0.512 | 0.040 | 12.86 | <2e−16 | *** |
| ESCO2_91 | 0.055 | 0.040 | 1.38 | 0.1691 | |
| ESCO2_92 | −2.551 | 0.040 | −64.05 | <2e−16 | *** |
| ESCO2_93 | 2.052 | 0.040 | 51.53 | <2e−16 | *** |
| ESCO2_94 | −0.742 | 0.040 | −18.64 | <2e−16 | *** |
| ESCO2_95 | −4.638 | 0.042 | −111.31 | <2e−16 | *** |
| ESCO2_96 | −0.940 | 0.040 | −23.60 | <2e−16 | *** |
| Nuts2_DE12 | −0.297 | 0.039 | −7.56 | 0.0000 | *** |
| Nuts2_DE13 | −0.893 | 0.039 | −22.72 | <2e−16 | *** |
| Nuts2_DE14 | −0.944 | 0.039 | −23.99 | <2e−16 | *** |
| Nuts2_DE21 | 0.242 | 0.039 | 6.16 | 0.0000 | *** |
| Nuts2_DE22 | −2.192 | 0.040 | −55.35 | <2e−16 | *** |
| Nuts2_DE23 | −1.822 | 0.039 | −46.18 | <2e−16 | *** |
| Nuts2_DE24 | −2.081 | 0.039 | −52.93 | <2e−16 | *** |
| Nuts2_DE25 | −1.130 | 0.039 | −28.75 | <2e−16 | *** |
| Nuts2_DE26 | −1.654 | 0.039 | −42.08 | <2e−16 | *** |
| Nuts2_DE27 | −1.502 | 0.039 | −38.07 | <2e−16 | *** |
| Nuts2_DE30 | 0.365 | 0.039 | 9.27 | <2e−16 | *** |
| Nuts2_DE40 | −0.627 | 0.039 | −15.94 | <2e−16 | *** |
| Nuts2_DE50 | −1.287 | 0.039 | −32.74 | <2e−16 | *** |
| Nuts2_DE60 | −0.184 | 0.039 | −4.68 | 0.0000 | *** |
| Nuts2_DE71 | −0.373 | 0.039 | −9.48 | <2e−16 | *** |
| Nuts2_DE72 | −2.179 | 0.040 | −55.07 | <2e−16 | *** |
| Nuts2_DE73 | −1.702 | 0.039 | −43.29 | <2e−16 | *** |
| Nuts2_DE80 | −1.170 | 0.039 | −29.77 | <2e−16 | *** |
| Nuts2_DE91 | −1.161 | 0.039 | −29.42 | <2e−16 | *** |
| Nuts2_DE92 | −0.830 | 0.039 | −21.12 | <2e−16 | *** |
| Nuts2_DE93 | −1.499 | 0.039 | −37.99 | <2e−16 | *** |
| Nuts2_DE94 | −1.019 | 0.039 | −25.88 | <2e−16 | *** |
| Nuts2_DEA1 | 0.135 | 0.039 | 3.42 | 0.0006 | *** |
| Nuts2_DEA2 | −0.183 | 0.039 | −4.64 | 0.0000 | *** |
| Nuts2_DEA3 | −0.867 | 0.039 | −21.98 | <2e−16 | *** |
| Nuts2_DEA4 | −0.933 | 0.039 | −23.70 | <2e−16 | *** |
| Nuts2_DEA5 | −0.641 | 0.039 | −16.31 | <2e−16 | *** |
| Nuts2_DEB1 | −1.404 | 0.039 | −35.72 | <2e−16 | *** |
| Nuts2_DEB2 | −2.640 | 0.039 | −66.94 | <2e−16 | *** |
| Nuts2_DEB3 | −0.570 | 0.039 | −14.49 | <2e−16 | *** |
| Nuts2_DEC0 | −1.680 | 0.039 | −42.66 | <2e−16 | *** |
| Nuts2_DED2 | −0.874 | 0.039 | −22.20 | <2e−16 | *** |
| Nuts2_DED4 | −1.421 | 0.039 | −36.16 | <2e−16 | *** |
| Nuts2_DED5 | −0.988 | 0.039 | −25.12 | <2e−16 | *** |

| OJA equation (logOJA-Gaussian Margin) | | | | | |
|---|---|---|---|---|---|
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | **Sign** |
| Nuts2_DEE0 | −0.869 | 0.039 | −22.06 | <2e−16 | *** |
| Nuts2_DEF0 | −0.464 | 0.039 | −11.80 | <2e−16 | *** |
| Nuts2_DEG0 | −0.712 | 0.039 | −18.12 | <2e−16 | *** |
| Quarter22_2 | 0.112 | 0.013 | 8.80 | <2e−16 | *** |
| Quarter22_3 | 0.288 | 0.013 | 22.55 | <2e−16 | *** |
| Quarter22_4 | 0.201 | 0.013 | 15.70 | <2e−16 | *** |
| Tech_EMPL_J | 2.051 | 0.040 | 51.53 | <2e−16 | *** |
| Tech_EMPL_M | 4.638 | 0.042 | 111.31 | <2e−16 | *** |

*Note:* Signif. codes: "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1.

**France**

Residual rho (95% Conf.Int.) = −0.0543(−0.283, 0.146).

| Selection equation (Probit margin) | | | | | |
|---|---|---|---|---|---|
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | |
| Intercept | −1.634 | 0.942 | −1.74 | 0.0827 | . |
| Time | 0.138 | 0.073 | 1.89 | 0.0586 | . |
| Tech_EMPL_N | 0.359 | 0.142 | 2.52 | 0.0116 | * |
| HomeBank21 | 0.080 | 0.012 | 6.65 | 0.0000 | *** |
| Profiles′dummies (Nuts1 × ESCO2) Chi-sq. | | | 148.00 | <2e−16 | *** |
| **OJA equation (OJA-Negative Binomial margin)** | | | | | |
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | |
| Intercept | 7.851 | 0.083 | 94.12 | <2e−16 | *** |
| ESCO2_12 | 1.521 | 0.100 | 15.20 | <2e−16 | *** |
| ESCO2_13 | 0.462 | 0.100 | 4.61 | 0.0000 | *** |
| ESCO2_14 | −0.359 | 0.100 | −3.58 | 0.0003 | *** |
| ESCO2_21 | 1.354 | 0.100 | 13.53 | <2e−16 | *** |
| ESCO2_22 | 0.746 | 0.100 | 7.45 | 0.0000 | *** |
| ESCO2_23 | −0.885 | 0.100 | −8.81 | <2e−16 | *** |
| ESCO2_24 | 1.822 | 0.100 | 18.21 | <2e−16 | *** |
| ESCO2_25 | 1.440 | 0.100 | 14.40 | <2e−16 | *** |
| ESCO2_26 | 0.702 | 0.100 | 7.01 | 0.0000 | *** |
| ESCO2_31 | 2.016 | 0.100 | 20.15 | <2e−16 | *** |
| ESCO2_32 | 0.702 | 0.100 | 7.01 | 0.0000 | *** |
| ESCO2_33 | 2.589 | 0.100 | 25.87 | <2e−16 | *** |
| ESCO2_34 | 0.880 | 0.100 | 8.79 | <2e−16 | *** |
| ESCO2_35 | −0.105 | 0.100 | −1.05 | 0.2951 | |
| ESCO2_41 | 0.190 | 0.100 | 1.90 | 0.0578 | . |
| ESCO2_42 | 0.096 | 0.100 | 0.96 | 0.3365 | |
| ESCO2_43 | 1.545 | 0.100 | 15.44 | <2e−16 | *** |
| ESCO2_44 | 1.761 | 0.100 | 17.60 | <2e−16 | *** |
| ESCO2_51 | 1.369 | 0.100 | 13.69 | <2e−16 | *** |

| | OJA equation (OJA-Negative Binomial margin) | | | | |
|---|---|---|---|---|---|
| | Estimate | Std.Error | *z* value | Pr(>\|*z*\|) | |
| ESCO2_52 | 1.966 | 0.100 | 19.65 | <2e−16 | *** |
| ESCO2_53 | 0.931 | 0.100 | 9.30 | <2e−16 | *** |
| ESCO2_54 | −0.535 | 0.100 | −5.33 | 0.0000 | *** |
| ESCO2_61 | −1.996 | 0.101 | −19.76 | <2e−16 | *** |
| ESCO2_62 | −5.170 | 0.115 | −45.13 | <2e−16 | *** |
| ESCO2_71 | 1.028 | 0.100 | 10.26 | <2e−16 | *** |
| ESCO2_72 | 1.327 | 0.100 | 13.26 | <2e−16 | *** |
| ESCO2_73 | −2.726 | 0.102 | −26.69 | <2e−16 | *** |
| ESCO2_74 | 0.663 | 0.100 | 6.62 | 0.0000 | *** |
| ESCO2_75 | 0.397 | 0.100 | 3.96 | 0.0001 | *** |
| ESCO2_81 | 0.827 | 0.100 | 8.25 | <2e−16 | *** |
| ESCO2_82 | −0.863 | 0.100 | −8.59 | <2e−16 | *** |
| ESCO2_83 | 1.322 | 0.100 | 13.20 | <2e−16 | *** |
| ESCO2_91 | 0.944 | 0.100 | 9.42 | <2e−16 | *** |
| ESCO2_92 | −2.226 | 0.101 | −21.99 | <2e−16 | *** |
| ESCO2_93 | 2.496 | 0.100 | 24.94 | <2e−16 | *** |
| ESCO2_94 | 0.153 | 0.100 | 1.53 | 0.1261 | |
| ESCO2_95 | −2.943 | 0.102 | −28.79 | <2e−16 | *** |
| ESCO2_96 | −1.822 | 0.101 | −18.06 | <2e−16 | *** |
| Nuts2_FRB0 | −1.373 | 0.063 | −21.82 | <2e−16 | *** |
| Nuts2_FRC1 | −1.538 | 0.063 | −24.43 | <2e−16 | *** |
| Nuts2_FRC2 | −1.767 | 0.063 | −28.07 | <2e−16 | *** |
| Nuts2_FRD1 | −1.891 | 0.063 | −29.98 | <2e−16 | *** |
| Nuts2_FRD2 | −1.556 | 0.063 | −24.69 | <2e−16 | *** |
| Nuts2_FRE1 | −1.163 | 0.063 | −18.49 | <2e−16 | *** |
| Nuts2_FRE2 | −1.651 | 0.063 | −26.21 | <2e−16 | *** |
| Nuts2_FRF1 | −1.476 | 0.063 | −23.45 | <2e−16 | *** |
| Nuts2_FRF2 | −2.090 | 0.063 | −33.13 | <2e−16 | *** |
| Nuts2_FRF3 | −1.504 | 0.063 | −23.90 | <2e−16 | *** |
| Nuts2_FRG0 | −0.970 | 0.063 | −15.43 | <2e−16 | *** |
| Nuts2_FRH0 | −1.115 | 0.063 | −17.72 | <2e−16 | *** |
| Nuts2_FRI1 | −0.937 | 0.063 | −14.91 | <2e−16 | *** |
| Nuts2_FRI2 | −2.960 | 0.063 | −46.75 | <2e−16 | *** |
| Nuts2_FRI3 | −1.502 | 0.063 | −23.87 | <2e−16 | *** |
| Nuts2_FRJ1 | −1.187 | 0.063 | −18.86 | <2e−16 | *** |
| Nuts2_FRJ2 | −0.981 | 0.063 | −15.61 | <2e−16 | *** |
| Nuts2_FRK1 | −1.674 | 0.063 | −26.60 | <2e−16 | *** |
| Nuts2_FRK2 | 0.016 | 0.063 | 0.25 | 0.8018 | |
| Nuts2_FRL0 | −0.769 | 0.063 | −12.23 | <2e−16 | *** |
| Nuts2_FRM0 | −2.723 | 0.063 | −42.89 | <2e−16 | *** |
| Nuts2_FRY1 | −5.799 | 0.067 | −86.06 | <2e−16 | *** |

| OJA equation (OJA-Negative Binomial margin) | | | | | |
|---|---|---|---|---|---|
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | |
| Nuts2_FRY2 | −5.444 | 0.066 | −81.87 | < 2e−16 | *** |
| Nuts2_FRY3 | −5.619 | 0.067 | −84.05 | < 2e−16 | *** |
| Nuts2_FRY4 | −1.437 | 0.063 | −22.80 | < 2e−16 | *** |
| Time | 0.054 | 0.004 | 13.94 | < 2e−16 | *** |
| Tech_EMPL_J | 2.226 | 0.101 | 21.99 | < 2e−16 | *** |
| Tech_EMPL_M | 2.592 | 0.101 | 25.87 | < 2e−16 | *** |
| Profiles'dummies (Nuts1 × ESCO2) Chi-sq. | | | 1614.10 | < 2e−16 | *** |

*Note:* Signif. codes: "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1.

**Italy**

Residual rho (95% Conf. Int.) = 0.0093(−0.147, 0.170).

| Selection equation (Probit Margin) | | | | | |
|---|---|---|---|---|---|
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | |
| Intercept | 2.080 | 0.119 | 17.48 | < 2e−16 | *** |
| Quarter22_2 | 0.037 | 0.147 | 0.25 | 0.8014 | |
| Quarter22_3 | 0.347 | 0.164 | 2.11 | 0.0349 | * |
| Quarter22_4 | 0.183 | 0.153 | 1.20 | 0.2323 | |
| Profiles'dummies (Nuts1xESCO2) Chi-sq. | | | 527.70 | < 2e−16 | *** |
| **OJA equation (logOJA-Gaussian margin)** | | | | | |
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | |
| Intercept | 5.774 | 0.063 | 91.94 | < 2e−16 | *** |
| ESCO2_12 | 1.510 | 0.069 | 21.89 | < 2e−16 | *** |
| ESCO2_13 | 0.505 | 0.069 | 7.32 | 0.0000 | *** |
| ESCO2_14 | 1.222 | 0.069 | 17.73 | < 2e−16 | *** |
| ESCO2_21 | 2.188 | 0.069 | 31.73 | < 2e−16 | *** |
| ESCO2_22 | 1.119 | 0.069 | 16.23 | < 2e−16 | *** |
| ESCO2_23 | 0.350 | 0.069 | 5.08 | 0.0000 | *** |
| ESCO2_24 | 2.186 | 0.069 | 31.71 | < 2e−16 | *** |
| ESCO2_25 | 2.211 | 0.069 | 32.07 | < 2e−16 | *** |
| ESCO2_26 | 0.775 | 0.069 | 11.24 | < 2e−16 | *** |
| ESCO2_31 | 1.982 | 0.069 | 28.74 | < 2e−16 | *** |
| ESCO2_32 | 0.458 | 0.069 | 6.64 | 0.0000 | *** |
| ESCO2_33 | 2.907 | 0.069 | 42.16 | < 2e−16 | *** |
| ESCO2_34 | 1.473 | 0.069 | 21.36 | < 2e−16 | *** |
| ESCO2_35 | 0.704 | 0.069 | 10.21 | < 2e−16 | *** |
| ESCO2_41 | 1.177 | 0.069 | 17.07 | < 2e−16 | *** |
| ESCO2_42 | 1.326 | 0.069 | 19.23 | < 2e−16 | *** |
| ESCO2_43 | 2.028 | 0.069 | 29.41 | < 2e−16 | *** |
| ESCO2_44 | 0.778 | 0.069 | 11.28 | < 2e−16 | *** |
| ESCO2_51 | 2.128 | 0.069 | 30.87 | < 2e−16 | *** |
| ESCO2_52 | 2.750 | 0.069 | 39.88 | < 2e−16 | *** |
| ESCO2_53 | 0.382 | 0.069 | 5.53 | 0.0000 | *** |

| | OJA equation (logOJA-Gaussian margin) | | | | |
|---|---|---|---|---|---|
| | Estimate | Std.Error | *z* value | Pr(>\|*z*\|) | |
| ESCO2_54 | −0.252 | 0.069 | −3.64 | 0.0003 | *** |
| ESCO2_61 | −4.536 | 0.085 | −53.67 | <2e−16 | *** |
| ESCO2_62 | −5.197 | 0.127 | −40.99 | <2e−16 | *** |
| ESCO2_71 | 0.629 | 0.069 | 9.13 | <2e−16 | *** |
| ESCO2_72 | 1.459 | 0.069 | 21.16 | <2e−16 | *** |
| ESCO2_73 | −1.616 | 0.070 | −22.97 | <2e−16 | *** |
| ESCO2_74 | 1.777 | 0.069 | 25.78 | <2e−16 | *** |
| ESCO2_75 | 0.819 | 0.069 | 11.88 | <2e−16 | *** |
| ESCO2_81 | 0.926 | 0.069 | 13.43 | <2e−16 | *** |
| ESCO2_82 | 0.891 | 0.069 | 12.92 | <2e−16 | *** |
| ESCO2_83 | 0.733 | 0.069 | 10.64 | <2e−16 | *** |
| ESCO2_91 | 1.459 | 0.069 | 21.16 | <2e−16 | *** |
| ESCO2_92 | −1.628 | 0.070 | −23.39 | <2e−16 | *** |
| ESCO2_93 | 2.481 | 0.069 | 35.99 | <2e−16 | *** |
| ESCO2_94 | 0.894 | 0.069 | 12.97 | <2e−16 | *** |
| ESCO2_95 | −3.934 | 0.080 | −49.24 | <2e−16 | *** |
| ESCO2_96 | −0.877 | 0.069 | −12.69 | <2e−16 | *** |
| Nuts2_ITC2 | −3.859 | 0.052 | −73.74 | <2e−16 | *** |
| Nuts2_ITC3 | −1.401 | 0.052 | −27.13 | <2e−16 | *** |
| Nuts2_ITC4 | 1.219 | 0.051 | 23.93 | <2e−16 | *** |
| Nuts2_ITF1 | −1.715 | 0.052 | −33.20 | <2e−16 | *** |
| Nuts2_ITF2 | −3.886 | 0.052 | −74.27 | <2e−16 | *** |
| Nuts2_ITF3 | −0.770 | 0.051 | −15.05 | <2e−16 | *** |
| Nuts2_ITF4 | −1.375 | 0.052 | −26.68 | <2e−16 | *** |
| Nuts2_ITF5 | −2.972 | 0.052 | −57.23 | <2e−16 | *** |
| Nuts2_ITF6 | −2.321 | 0.052 | −44.96 | <2e−16 | *** |
| Nuts2_ITG1 | −1.522 | 0.051 | −29.64 | <2e−16 | *** |
| Nuts2_ITG2 | −2.141 | 0.051 | −41.61 | <2e−16 | *** |
| Nuts2_ITH1 | −2.043 | 0.052 | −39.65 | <2e−16 | *** |
| Nuts2_ITH2 | −2.053 | 0.051 | −39.89 | <2e−16 | *** |
| Nuts2_ITH3 | 0.331 | 0.051 | 6.49 | 0.0000 | *** |
| Nuts2_ITH4 | −1.259 | 0.052 | −24.43 | <2e−16 | *** |
| Nuts2_ITH5 | 0.523 | 0.051 | 10.23 | <2e−16 | *** |
| Nuts2_ITI1 | −0.391 | 0.051 | −7.65 | 0.0000 | *** |
| Nuts2_ITI2 | −2.194 | 0.052 | −42.40 | <2e−16 | *** |
| Nuts2_ITI3 | −1.404 | 0.052 | −27.24 | <2e−16 | *** |
| Nuts2_ITI4 | −0.104 | 0.051 | −2.05 | 0.0406 | * |
| Time | 0.096 | 0.007 | 13.43 | <2e−16 | *** |
| Tech_EMPL_J | 2.053 | 0.051 | 39.89 | <2e−16 | *** |
| Tech_EMPL_C | 4.536 | 0.085 | 53.67 | <2e−16 | *** |

*Note:* Signif. codes: "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1.

**Spain**

Residual rho (95% Conf. Int.) = 0.0885(−0.052,0.239).

| | Selection equation (Probit margin) | | | | |
|---|---|---|---|---|---|
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | |
| Intercept | −3.720 | 0.795 | −4.68 | 0.0000 | *** |
| ESCO2_12 | 1.751 | 0.699 | 2.51 | 0.0122 | * |
| ESCO2_13 | 6.457 | 1292.141 | 0.01 | 0.9960 | |
| ESCO2_14 | 0.418 | 0.501 | 0.84 | 0.4038 | |
| ESCO2_21 | 6.616 | 1297.802 | 0.01 | 0.9959 | |
| ESCO2_22 | 6.646 | 1330.851 | 0.01 | 0.9960 | |
| ESCO2_23 | 6.562 | 1052.745 | 0.01 | 0.9950 | |
| ESCO2_24 | 6.569 | 1245.099 | 0.01 | 0.9958 | |
| ESCO2_25 | 6.647 | 806.059 | 0.01 | 0.9934 | |
| ESCO2_26 | 6.531 | 1244.652 | 0.01 | 0.9958 | |
| ESCO2_31 | 6.538 | 1154.633 | 0.01 | 0.9955 | |
| ESCO2_32 | 6.477 | 1253.835 | 0.01 | 0.9959 | |
| ESCO2_33 | 6.615 | 1234.028 | 0.01 | 0.9957 | |
| ESCO2_34 | 6.501 | 1329.446 | 0.01 | 0.9961 | |
| ESCO2_35 | 0.206 | 0.490 | 0.42 | 0.6745 | |
| ESCO2_41 | 1.737 | 0.664 | 2.62 | 0.0089 | ** |
| ESCO2_42 | 1.670 | 0.657 | 2.54 | 0.0111 | * |
| ESCO2_43 | 0.426 | 0.507 | 0.84 | 0.4007 | |
| ESCO2_44 | −0.059 | 0.457 | −0.13 | 0.8980 | |
| ESCO2_51 | 6.629 | 1273.223 | 0.01 | 0.9959 | |
| ESCO2_52 | 6.551 | 1237.447 | 0.01 | 0.9958 | |
| ESCO2_53 | 6.595 | 1289.528 | 0.01 | 0.9959 | |
| ESCO2_54 | 1.153 | 0.587 | 1.96 | 0.0496 | * |
| ESCO2_61 | 0.375 | 0.505 | 0.74 | 0.4571 | |
| ESCO2_62 | −3.151 | 0.408 | −7.73 | 0.0000 | *** |
| ESCO2_71 | 6.591 | 1332.763 | 0.01 | 0.9961 | |
| ESCO2_72 | 6.729 | 1399.466 | 0.01 | 0.9962 | |
| ESCO2_73 | 0.181 | 0.486 | 0.37 | 0.7095 | |
| ESCO2_74 | 1.324 | 0.590 | 2.25 | 0.0248 | * |
| ESCO2_75 | 0.641 | 0.526 | 1.22 | 0.2232 | |
| ESCO2_81 | 1.884 | 0.681 | 2.77 | 0.0057 | ** |
| ESCO2_82 | 6.625 | 1240.738 | 0.01 | 0.9957 | |
| ESCO2_83 | 0.978 | 0.555 | 1.76 | 0.0779 | . |
| ESCO2_91 | 1.269 | 0.587 | 2.16 | 0.0307 | * |
| ESCO2_92 | −0.553 | 0.439 | −1.26 | 0.2079 | |
| ESCO2_93 | 1.275 | 0.578 | 2.21 | 0.0273 | * |
| ESCO2_94 | 0.014 | 0.480 | 0.03 | 0.9766 | |
| ESCO2_95 | −2.912 | 0.401 | −7.27 | 0.0000 | *** |
| ESCO2_96 | −0.648 | 0.434 | −1.49 | 0.1354 | |

| Selection equation (Probit margin) | | | | | |
|---|---|---|---|---|---|
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | |
| Nuts1ES2 | 2.464 | 0.447 | 5.51 | 0.0000 | *** |
| Nuts1ES3 | 8.550 | 1368.500 | 0.01 | 0.9950 | |
| Nuts1ES4 | −0.090 | 0.286 | −0.32 | 0.7524 | |
| Nuts1ES5 | 1.335 | 0.334 | 4.00 | 0.0001 | *** |
| Nuts1ES6 | −2.906 | 0.303 | −9.58 | < 2e−16 | *** |
| Nuts1ES7 | −0.803 | 0.483 | −1.66 | 0.0967 | . |
| Quarter22_2 | 0.355 | 0.192 | 1.85 | 0.0643 | . |
| Quarter22_3 | 0.554 | 0.208 | 2.67 | 0.0076 | ** |
| Quarter22_4 | 0.479 | 0.194 | 2.46 | 0.0138 | * |
| WEBO21_A_B | 0.350 | 0.043 | 8.20 | 0.0000 | *** |
| WEBO21_G_I_T | 0.148 | 0.021 | 6.92 | 0.0000 | *** |
| WEBO21_O_U | 0.067 | 0.017 | 4.06 | 0.0000 | *** |
| **OJA equation (logOJA-Gaussian margin)** | | | | | |
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | |
| Intercept | 2.701 | 0.078 | 34.51 | < 2e−16 | *** |
| ESCO2_12 | 2.560 | 0.087 | 29.49 | < 2e−16 | *** |
| ESCO2_13 | 1.899 | 0.087 | 21.93 | < 2e−16 | *** |
| ESCO2_14 | 1.030 | 0.088 | 11.71 | < 2e−16 | *** |
| ESCO2_21 | 3.470 | 0.087 | 40.07 | < 2e−16 | *** |
| ESCO2_22 | 2.738 | 0.087 | 31.61 | < 2e−16 | *** |
| ESCO2_23 | 3.049 | 0.087 | 35.21 | < 2e−16 | *** |
| ESCO2_24 | 3.031 | 0.087 | 35.00 | < 2e−16 | *** |
| ESCO2_25 | 3.311 | 0.087 | 38.24 | < 2e−16 | *** |
| ESCO2_26 | 2.529 | 0.087 | 29.20 | < 2e−16 | *** |
| ESCO2_31 | 3.153 | 0.087 | 36.41 | < 2e−16 | *** |
| ESCO2_32 | 2.156 | 0.087 | 24.89 | < 2e−16 | *** |
| ESCO2_33 | 4.112 | 0.087 | 47.49 | < 2e−16 | *** |
| ESCO2_34 | 2.390 | 0.087 | 27.59 | < 2e−16 | *** |
| ESCO2_35 | 1.941 | 0.088 | 22.00 | < 2e−16 | *** |
| ESCO2_41 | 2.309 | 0.087 | 26.59 | < 2e−16 | *** |
| ESCO2_42 | 2.167 | 0.087 | 24.95 | < 2e−16 | *** |
| ESCO2_43 | 1.687 | 0.088 | 19.19 | < 2e−16 | *** |
| ESCO2_44 | 1.190 | 0.089 | 13.44 | < 2e−16 | *** |
| ESCO2_51 | 3.590 | 0.087 | 41.46 | < 2e−16 | *** |
| ESCO2_52 | 3.647 | 0.087 | 42.11 | < 2e−16 | *** |
| ESCO2_53 | 2.420 | 0.087 | 27.95 | < 2e−16 | *** |
| ESCO2_54 | 0.884 | 0.087 | 10.15 | < 2e−16 | *** |
| ESCO2_61 | 0.423 | 0.088 | 4.81 | 0.0000 | *** |
| ESCO2_62 | −2.888 | 0.130 | −22.14 | < 2e−16 | *** |
| ESCO2_71 | 2.669 | 0.087 | 30.82 | < 2e−16 | *** |
| ESCO2_72 | 2.841 | 0.087 | 32.81 | < 2e−16 | *** |

| | OJA equation (logOJA-Gaussian margin) | | | | |
|---|---|---|---|---|---|
| | **Estimate** | **Std.Error** | ***z* value** | **Pr(>\|*z*\|)** | |
| ESCO2_73 | −0.247 | 0.088 | −2.80 | 0.0052 | ** |
| ESCO2_74 | 2.826 | 0.087 | 32.45 | <2e−16 | *** |
| ESCO2_75 | 1.857 | 0.088 | 21.18 | <2e−16 | *** |
| ESCO2_81 | 3.041 | 0.087 | 35.03 | <2e−16 | *** |
| ESCO2_82 | 1.805 | 0.087 | 20.85 | <2e−16 | *** |
| ESCO2_83 | 2.567 | 0.087 | 29.39 | <2e−16 | *** |
| ESCO2_91 | 2.158 | 0.087 | 24.78 | <2e−16 | *** |
| ESCO2_92 | −0.851 | 0.090 | −9.50 | <2e−16 | *** |
| ESCO2_93 | 2.196 | 0.087 | 25.21 | <2e−16 | *** |
| ESCO2_94 | 1.925 | 0.089 | 21.74 | <2e−16 | *** |
| ESCO2_95 | −2.005 | 0.124 | −16.23 | <2e−16 | *** |
| ESCO2_96 | −0.288 | 0.090 | −3.21 | 0.0014 | ** |
| Nuts2_ES12 | −0.796 | 0.060 | −13.21 | <2e−16 | *** |
| Nuts2_ES13 | −1.333 | 0.060 | −22.11 | <2e−16 | *** |
| Nuts2_ES21 | 0.287 | 0.060 | 4.78 | 0.0000 | *** |
| Nuts2_ES22 | −0.792 | 0.060 | −13.23 | <2e−16 | *** |
| Nuts2_ES23 | −1.714 | 0.061 | −28.28 | <2e−16 | *** |
| Nuts2_ES24 | −0.100 | 0.060 | −1.67 | 0.0951 | . |
| Nuts2_ES30 | 1.621 | 0.060 | 27.15 | <2e−16 | *** |
| Nuts2_ES41 | 0.038 | 0.060 | 0.63 | 0.5275 | |
| Nuts2_ES42 | −0.328 | 0.060 | −5.44 | 0.0000 | *** |
| Nuts2_ES43 | −1.069 | 0.060 | −17.69 | <2e−16 | *** |
| Nuts2_ES51 | 1.788 | 0.060 | 29.95 | <2e−16 | *** |
| Nuts2_ES52 | 0.532 | 0.060 | 8.87 | <2e−16 | *** |
| Nuts2_ES53 | −0.732 | 0.060 | −12.14 | <2e−16 | *** |
| Nuts2_ES61 | 0.890 | 0.060 | 14.87 | <2e−16 | *** |
| Nuts2_ES62 | −0.785 | 0.060 | −13.00 | <2e−16 | *** |
| Nuts2_ES63 | −4.301 | 0.066 | −65.09 | <2e−16 | *** |
| Nuts2_ES64 | −4.210 | 0.067 | −63.00 | <2e−16 | *** |
| Nuts2_ES70 | −0.441 | 0.060 | −7.32 | 0.0000 | *** |
| Time | 0.098 | 0.009 | 11.04 | <2e−16 | *** |
| Tech_EMPL_C | 3.040 | 0.089 | 37.02 | <2e−16 | *** |
| Tech_EMPL_M | 2.158 | 0.087 | 24.78 | <2e−16 | *** |

*Note:* Signif. codes: "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1.