Journal of Forecasting

WILEY

RESEARCH ARTICLE OPEN ACCESS

Forecasting New Employment Using Nonrepresentative Online Job Advertisements With an Application to the Italian and EU Labor Market

Pietro Giorgio Lovaglio[1,2] | Mario Mezzanzanica[1,2]

[1]Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy | [2]CRISP Research Center, University of Milano-Bicocca, Milan, Italy

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

Received: 25 February 2025 | **Revised:** 27 November 2025 | **Accepted:** 15 December 2025

Keywords: forecast | labor force survey | multilevel modeling | online job advertisements | poststratification | sample selection models

ABSTRACT

Using online job advertisement data improves the timeliness and granularity depth of analysis in the labor market in domains not covered by official data. Specifically, its variation over time may be used as an anticipator of official employment variations. However, online job advertisements may not be representative in terms of key labor market variables. The paper presents a methodology that forecasts publicly available official employment LFS recent job starters exploiting its relationship with a bias-corrected version of online job postings, obtained by predictions of a bivariate sample selection model, which jointly estimates the number of vacancies within job profiles and the probability of endogenous selection for nonzero vacancies. LFS new hires (obtained from LFS microdata) were used as benchmark data to measure the bias of online data and adjusted predicted counts. The proposed framework is illustrated using a dataset of Italian online job advertisements spanning from the period 2013-Q2 to 2018-Q2 to forecast quarterly LFS recent job starters 1 year ahead and the Cedefop's Skills-OVATE data using Italy, France, Germany, and Spain in 2022. Results demonstrated that raw vacancies present a strong bias level with respect to benchmark data, whereas sample selection models reduced this bias by half, unlike multilevel estimates. Moreover, LFS forecasts using a VECM that leverages cointegration between LFS recent job starters and adjusted online vacancy series offer a valuable alternative to traditional univariate forecasting methods.

JEL Classification: C13, J21, J23

1 | Introduction

Government policy has placed increasing emphasis on the need for robust labor market stock projections to assist in policy and planning for the provision of education and training.

Among various statistics, recently, Eurostat has begun reporting the number of recent job starters (individuals who began their employment within the 3 months preceding the interview) for each EU country and sector of economic activity (Eurostat 2024). This source, referred to in the paper as "LFS recent starters", is a quarterly series including (at the moment of writing this paper) the first quarter of 2025.

Forecasting such series is particularly appealing because, despite their excessively large aggregate stock, they benefit from near real-time dissemination, with only a one-quarter delay. Many studies that focus on labor outcomes from the supply side, such as employment or wages, often suggest that demand-side factors may be crucial determinants of employment shifts (and forecasts), even if these demand-side factors are not directly specified and measured in the analysis (Hamermesh 1993;

Card and Krueger 1994). Examples are employment variation (Rosen and Welch 1971; Moffitt 1984; Card and Krueger 1994; Hamermesh 1993) and the role of wage elasticities on employment variation (Chetty et al. 2011; Lichter et al. 2015; Neumark and Wascher 2007).

Moreover, specifying demand-side factors in labor market analyses is crucial because skill composition and shifts in skill mixes may not be adequately captured when focusing solely on employment-based approaches (Salvatori 2018; Caines et al. 2017; Deming 2017; Deming and Kahn 2018)

Recent EU labor market policies indicate that job vacancy data can be exploited to improve the knowledge and functioning of the labor market by matching supply and demand more closely (Eurostat 2023, OECD 2023; Cedefop 2019).

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; Couper 2013; Tam and Clarke 2015; Lovaglio et al. 2020; Hershbein and Kahn 2018; Zilian et al. 2021). In the same end, 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) agree with the potentiality of such data source for labor market research.

Recently, OJA data have been increasingly explored in connection with economic theories such as the Beveridge curve (Turrell et al. 2018) and the Phillips curve (Faryna et al. 2022), to identify emerging employment trends (Lovaglio 2022) and to measure skill mismatches and labor shortages (Cedefop 2023, 2024).

Previous studies examine job vacancy data as a possible leading indicator of cyclical employment dynamics (Haggar-Guénette 1989; Zagorsky 1998; Amoah 2000; Valletta 2005; Ruth et al. 2006; Mandrone et al. 2010), where lagged OJA growth anticipates employment growth (European Central Bank 2002; Australian Bureau of Statistics 2003; Lovaglio et al. 2020).

However, one less explored issue in the literature is that OJA may not fully represent the general (labor market) population (Fan et al. 2014; Fan and Liao 2012; Gelman et al. 2016; Tam and Clarke 2015; Kruskal and Mosteller 1979)

Broadly speaking, OJA data can suffer from representativeness issues, potentially favoring—for unknown reasons—certain geographical areas, sectors, or occupations more inclined to appear on web portals or to advertise open positions online. In this end, a reference sample or benchmark data from census or official surveys, fully representative at the population level, is strongly advocated.

Supposing to analyze a continuous outcome y for n observations, where the objective is to estimate a parameter θ (the total of OJA in the population). When a nonrepresentative sample is obtained, estimates of θ are biased. Nonrepresentativeness is often defined (Kruskal and Mosteller 1979) in terms of differences in the distributions of y between the sample and the population (or a benchmark representative sample) within levels of relevant auxiliary variables (e.g., sector and region).

The issue of representativeness is linked to the presence of a nonrandom (not ignorable) missing data process (Kreuter and Peng 2014), which arises from nonrandom 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)

The missing data process arises when the sample does not cover all levels of the population, that is, when certain profiles (combinations of auxiliary variables) are absent from the nonrepresentative sample—for example, missing OJA in specific sector–region cells.

OJA data are prone to nonrandom selectivity because platform collection procedures are not designed for statistical purposes. Observed samples of online vacancies are likely affected by a nonrandom mechanism, such as incomplete coverage of online job advertisements, selective website inclusion, and underrepresentation/overrepresentation of certain sectors, regions, and occupations. Consequently, this selectivity leads to coverage gaps and nonresponse (or missing data), potentially biasing OJA data and forecasts of related series, such as employment.

This paper primarily focuses on forecasting the official employment Labour Force Survey (LFS) Recent Job Starters series using nonrepresentative OJA series as auxiliary data in bivariate forecasting models.

Specifically, by employing a methodology that accounts for the nonrandom selectivity of OJA, we obtained a bias-corrected version of vacancy data. These adjusted predictions are more representative of the benchmark and can be used to forecast LFS recent starters within bivariate forecasting frameworks, leveraging shared time series properties such as common trends and cointegration.

The assessment of representativeness (and potential bias) was performed by comparing the distributions of the models' predictions of different approaches and raw OJA with the reference sample (new hires obtained from LFS microdata) used as benchmark data.

Among different approaches, one well-established strategy for handling nonrandom selectivity in data is known in the literature as multilevel regression and poststratification (MRP, Gelman and Little 1997; Gelman et al. 2016; Wang et al. 2014), a two-step strategy using multilevel to find predictions and individual data and then recalibrating (poststratification) these estimates in line with an established reference sample.

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, 2012; Warshaw and Rodden 2012;

Buttice and Highton 2013; Wang et al. 2014; Gelman et al. 2007, 2009).

However, this approach works essentially with a missing at random (MAR) mechanism, assuming that all relevant sources (covariates and auxiliary variables) underlying the missingness mechanism have been specified in the multilevel model and in the poststratification phase (Pfeffermann 2007; Little 2007).

Various strategies have been proposed in the literature to address the missing not at random (MNAR) mechanism in different contexts. Among others, sample selection models (SSMs, Heckman 1976; Maddala 1983) jointly model both the outcome equation (counts of) and the probability that an observation is missing or included in the sample (selection equation).

Particularly, generalized joint regression models (GJRMs; Lee 1983; Smith 2003; Winkelmann 2011; Marra and Radice 2013; Radice et al. 2016) provide consistent estimators for joint outcome and selection models, extending classical SSM to a broader range of empirical scenarios by accommodating diverse error distributions and dependencies between equations.

In the present paper, we propose an SSM strategy based on GJRM to address and correct nonrandom selectivity, facilitating the development of more accurate and robust measures of vacancy distributions to be used in the forecast strategy.

As illustrative examples, we apply this methodology to an official dataset of Italian OJA from the period 2013-Q2 to 2018-Q2 and to more recent data from 2022 for the four most populous EU countries, using Cedefop's Skills-OVATE dataset.

To our knowledge, this is the first paper to measure the bias of two OJA data sources (Italian and European) against LFS benchmark data (both aggregated and microdata) and to propose an SSM as a strategy to mitigate such bias.

Results demonstrated that for both Italian and EU data, raw vacancies present a strong bias level with respect to benchmark LFS data, whereas predictions from SSM/GJRM reduced this bias by half, unlikely multilevel predictions that resulted strongly biased.

Empirically, unlike previous results that found strong relationships among variables in differences, we found a significant relationship among levels of adjusted OJA and employment series: particularly, these bias-corrected predictions of vacancy data resulted in cointegration with LFS recent starts. To this end, a bivariate forecasting model exploiting the structural long-run relationship among two series strongly overperformed traditional univariate forecasting methods in terms of forecast accuracy.

The paper is organized as follows: Section 2 briefly reviews the statistical approaches used to work with nonrepresentative samples and possible benchmark data for OJA. Section 3 outlines the methodological approach. Section 4 details the data utilized, whereas Section 5 presents the results. Section 6 discusses the main findings and outlines future challenges. Section 7 provides the conclusion.

2 | Dealing With Nonrepresentative Data

2.1 | Population Frame and OJA Benchmark Data

MRP is a perfect synthesis addressing nonrepresentative samples combining a superpopulation model-based and a pseudo-randomization or reweighting approach (Elliott and Valliant 2017; Valliant 2019; Buelens et al. 2018).

To estimate a population parameter (θ, e.g., a probability or a total count), MRP is a two-step strategy involving, first, a classical multilevel equation that models an outcome variable y_i using individual data ($i=1, \ldots, n$) with fixed covariates and random effects, exploiting the possible hierarchical data structure (e.g., individuals within regions within states or repeated measures for individuals) and, second, poststratification. This second step recalibrates the individual estimates (predictions) by the counts of a reference sample or a benchmark data, generally derived from census or register data, containing K categorical auxiliary variables, such as demographic and geographic characteristics the resume the characteristics of the population.

The population, without loss of generality, contains K categorical variables of interest (auxiliary variables), and the kth has J_k categories. 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 to the so-called *population frame* (P_J), also known as the poststratification frame.

Hence, each profile (row j) of the population frame contains individual data aggregated at level j ($j=1, \ldots, J$, where i belongs to j), and the set of J rows represents all possible combinations (profiles) of the K auxiliary variables in the population.

Each jth population profile has N_j observations. For example, in the labor market context where the aim is to estimate the employment rate (θ) using a nonprobability sample (S_J), P_J contains the reference population for all possible combinations of Age_class × Gender × Region × Quarter, available from official sources.

Moreover, using a nonprobability sample, the individual multilevel estimates of the parameter defining employment probability ($\hat{y}_i$) are aggregated at the same level (j) of the population frame ($\hat{y}_j$) and weighted by N_j. In this end, MRP estimates the parameter, adjusting the estimates by poststratification ($\hat{y}^{PS}=\sum_j \hat{y}_j N_j / \sum_j N_j$), correcting $\hat{y}_j$ by the known shares (N_j / N) of people in the population strata (Little 1993).

When dealing with OJA, when the aim is to estimate the amount of OJA (θ) over covariates' levels (profiles) of interest, individual OJA data can be reorganized at an aggregated level, exactly as described above to find the population frame.

This leads to the *sample frame* S_J, where each jth profile has y_j observations (the outcome variable defining the number of OJA

in the jth profile). In this context of OJA aggregated data, aggregated estimates $\hat{y}_j$ were directly produced by the multilevel equation (first MRP step).

2.2 | Benchmark Data for OJA

Given the absence of data on the full OJA population (population frame), other official sources must be used as representative benchmarks for the OJA sample data.

In the general framework of analyzing the labor market and specifically OJA in Europe, Eurostat data are an imperative source (Beręsewicz et al. 2018; Zilian et al. 2021; Cedefop 2023, 2024). 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 and France) without specifying the actual number of positions demanded. Additionally, JVS does not encompass all sectors, as the covered sectors vary by country, and it offers limited detail at the occupational or regional levels, with a few exceptions. For example, Belgium, Denmark, Hungary, and Romania provide vacancy statistics by one-digit NUTS region, whereas Hungary and North Macedonia also report by one-digit ISCO occupation.

Moreover, as JVS data result from separate national surveys on vacancies, they are collected using varying methods and definitions across EU countries, including differing definitions of a quarter associated with a job vacancy. Consequently, these data offer a limited assessment of the underlying conditions of the labor market in real time.

Another essential source is LFS (Eurostat 2021) containing both microdata and aggregated data that refer to labor supply. One suggested strategy is to compare OJA data with LFS employment data (Hershbein and Kahn 2018; de Pedraza et al. 2019; Lovaglio 2022), specifically focusing on new jobs created in the past 3 months for each quarter, as suggested by recent literature (Lovaglio et al. 2020; Garasto et al. 2021; Cammeraat and Squicciarini 2021; Turrell et al. 2018).

To this end, an ad hoc benchmark dataset can be constructed from the richness of LFS microdata, allowing for systematic comparison of OJA (and predicted OJA values generated by different statistical approaches) with official data.

In particular, we can select a subset of new hires (LFS new hires) that aligns the above mentioned LFS recent job starters more closely with our objectives, as an example, focusing solely on employees and excluding unpaid workers or self-employed or new hires in public administration. Moreover, the LFS microdata offer a very granular population frame, including detailed variables such as ESCO2 (Occupations) × Nuts2 (Region) × Nace (Sector) × Quarter.

Despite the delay in microdata dissemination that prevents their use for near-real-time forecasts in conjunction with OJA, they can be used to assess the bias of OJA predictions relative to an unbiased official source across different statistical approaches when fitted on historical data. This is how we utilize LFS microdata in the paper.

2.3 | Missingness

The previous discussion has motivated the use of LFS's new hires as a population frame. However, although covariates that define S_J are the same as those that define P_J, the sample frame (OJA) does not necessarily encompass all the J levels of the population frame, particularly when certain profiles (combinations of auxiliary variables) are not represented in the sample or when the outcome is missing for such profiles. This leads to a problem of missing data (absence of OJA) in certain profiles of S_J. If these missing data are not at random, S_J resulted in a truncated (nonrepresentative) sample from a more general overall population.

Apart from some exceptions (Matei 2018; Elliott and Valliant 2017; Zhang et al. 2013), model-based approaches, and thus also MRP, rely on the assumption that missing data are MAR, that is, conditioning on individual covariates and on auxiliary variables (that define the population profiles); nonresponse is independent of the (unobserved) outcome variable.

In general, 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).

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 vein, 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."

If this does not hold, the observed portion of the population differs from the unobserved portion, and this constitutes a standard case of nonrandom sample truncation (Heckman 1976; Heckman 1979), or nonrandom 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 an MNAR situation.

In this end, another limitation of MRP is that, including another source of erratic variability (the random effects), it relies on a broad range of assumptions about error independence and exogeneity of covariates (Gelman 2014), many of which are generally not testable.

This rigidity is further compounded in possible situations of nonrandom selectivity, where the intricate nature of the selection mechanism complicates the analysis even when only a single source of error is present. As a 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.

MGCV specifies the same random effects of the multilevel equation, not precisely as random variables, but as penalized fixed effects, whose coefficients were estimated exactly as random effects (Robertson 1955; Wood 2004), thus exploiting the hierarchy in the data as multilevel, also avoiding complications by the specification of a new erratic source of variability.

Various strategies have been proposed in the literature to address the 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 SSMs à la Heckman (Heckman 1976; Maddala 1983; Valliant et al. 2000; Terza 1998). These methods can adjust for biases that may arise from a possible MNAR mechanism.

SSMs jointly model both the outcome equation (counts of) 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.

One limitation of the classical SSM approaches is that they jointly model two endogenous outcomes under the hypothesis that both errors follow a bivariate Gaussian distribution, which can be restrictive in many empirical applications. When this assumption is violated—such as in cases of skewed or heteroskedastic errors—the model's estimates may become inconsistent and biased (Vella 1998; Newey et al. 1990).

GJRMs (Smith 2003; Winkelmann 2011; Marra and Radice 2013) represent viable and robust alternatives, because they provide consistent estimators of a joint (outcome and selection) model with continuous or discrete (counts) outcomes while also enabling the specification of effects to leverage the hierarchical structure of (longitudinal) data. This is the methodology we employ, as detailed in the next section.

3 | Proposed Methodology

In this section, we describe how to obtain an adjusted version of OJA data and strategies to forecast LFS (recent starters).

Regarding the first issue, firstly, we obtain the sample frame S_J, as above described *but including* all possible population levels (profiles or cells) derived from a *sparse* cross-classification of the K auxiliary covariates. The sparse approach ensures that all existing profiles in the population appear in the sample frame S_J, including those levels not present in the sample data due to missing levels of some covariates in the sampled units.

In our context, a researcher generally works with OJA counts across the levels of the most relevant labor market variables, such as R regions, S sectors, O occupations, and T quarters. S_J thus may be defined by $R \times S \times O \times T$ profiles (rows).

For models that manage 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 jth 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 r, s, and o represent all possible levels of these variables from official sources (such as Nuts2, Nace, and ESCO2, respectively), regardless of whether OJA is present in the sample data.

Each row of S_J represents the new jth observation unit, and each level (profile) contains the outcome variable y_{jt}. The sparse S_J thus allows for two possible values for y_{jt} (OJA_{jt}, OJA counts in the jth profile at time t): either positive or missing, thus replicating the classical observation mechanism of the SSM framework.

On the sparse S_J, we can specify our SSM as a GJRM model, composed of two equations and a general form for the error structure, such as

$$z_{jt}^{*} = \alpha_0 + \alpha^T \mathbf{x}_{1jt} + u_{jt}, \tag{1}$$

$$y_{jt} = \beta_0 + \beta^T \mathbf{x}_{2jt} + e_{jt} \quad z_{jt=1}, \tag{2}$$

$$F(e_{jt}, u_{jt}) = C\left(F_1(e_{jt}), F_2(u_{jt}), \rho\right), \tag{3}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors of covariates of two equations (that may share some variables) and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the corresponding fixed parameters, and α_0 and β_0 the intercepts, respectively.

Particularly, observation rules are driven by two equations: a selection Equation (1) that models the determinants of missingness and an outcome Equation (2) that models the intensity of the outcome ($y_{jt} = \text{OJA}_{jt}$) in nonmissing profiles of the sample frame.

Regarding the selection equation, using the latent variable representation, z_{jt}^{*} represents the attitude that an OJA in profile j at time t is selected, the *selection* mechanism is thus represented by the Bernoulli variable z_{jt}, that equals $z_{jt} = 1$ (when $z_{jt}^{*} > 0$) for $\text{OJA}_{jt} > 0$ and 0 otherwise: y_{jt} can be observed only when $z_{jt} = 1$; otherwise, ($z_{jt} = 0$) y_{jt} is missing.

Regarding errors that are not profile-specific, $\boldsymbol{F}(e_{jt}$ and $u_{jt})$ is the bivariate cumulative distribution function (cdf) of error terms (u_{jt} and e_{jt}) expressed as a function $\boldsymbol{C}$, called copula, of one-dimensional margins $F_1(.)$ and $F_2(.)$ and their correlation ρ.

Equation (3) is the crucial aspect of GJRM flexibility. Broadly speaking, a copula is a function that connects 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 $F_1(y_{1j})$ and $F_2(y_{2j})$ for observation j, then there exists a two-dimensional copula $\boldsymbol{C}$ such that $\boldsymbol{F}(y_{1j}, y_{2j}) = \boldsymbol{C}(F_1(y_{1j}), F_2(y_{2j}); \rho_j)$, where y_1 and y_2 are two random variables and ρ_j represent an association parameter linking the correlation of two margins (Trivedi and Zimmer 2007).

Hence, the formulation $\boldsymbol{F}(e_{jt}, u_{jt}) = \boldsymbol{\Phi}_2(e_{jt}, u_{jt}, \rho)$ of Equation (3) is only a particular case among many possible specifications that can be expressed with copulas. In fact, using two Gaussian marginals $F_1(e_{jt})$ and $F_2(u_{jt})$ for error terms and a Gaussian copula (with fixed correlation ρ_j over J observations) $\boldsymbol{C} = \Phi_2(\Phi^{-1}(F_1), \Phi^{-1}(F_2); \rho)$, the copula representation of $\boldsymbol{F}(e_{jt}, u_{jt})$ leads the classical standard Heckman-type model: Equation (1) represents a Probit model, whereas Equation (2) is a regression model with Gaussian errors.

The more flexible copula version allows non-Gaussian dependence among error also leading to the strength and direction of the association ρ between the two marginals may vary across observations or groups of observations (ρ_j).

For the 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 in the $\mathbf{x}_2$ set. In our empirical context, webographics, such as internet penetration in societies and the regional labor market, were used as instruments.

The above equations 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, $\hat{y}_{jt}$) in case of nonrandom selectivity. Because MAR and MNAR mechanisms cannot be formally distinguished through a direct statistical test, we evaluate their plausibility using both the significance of the correlation parameter (ρ) and the likelihood ratio test of the null hypothesis $\rho = 0$. A significant correlation between the two equations (ρ) or a significant LR test indicates the presence of sample selection, implying that the missing data mechanism is nonignorable (MNAR) and that the two equations should therefore be jointly estimated. Thus, univariate approaches such as multilevel that ignore the selection equation are not appropriate.

Moreover, SSM (in our GJRM setting) generally models the unconditional expected values $\hat{y}_{jt} = E(y_{jt} | \mathbf{x}_{2j})$ in the outcome equation, thus predicting the outcome of interest for both selected (profiles with $\text{OJA}_{jt} > 0$) and unselected observations (profiles with missing OJA_{jt}), an unlikely univariate approach that predicts outcome only for nonmissing profiles ($z_{jt} = 1$). These predictions, estimated under a possible MNAR scenario that adjusts the expectation of the outcome variable to account for differences between the observed and unobserved data, can be considered an adjusted version of OJA that adjusts for nonrandom selectivity bias.

This adjustment ensures that the predicted outcome accounts for selection effects, systematically correcting for bias. This correction is mediated by the effects of covariates specified both in the selection and outcome equations. Specifically, for observations with a high probability of selection, the model prevents overestimation by adjusting for the overrepresentation of certain characteristics (covariates), whereas for observations with a low probability of selection, it mitigates underestimation by appropriately weighting their contribution to the estimation.

3.1 | Forecast

We forecast LFS recent starters at time t (LFS_t), exploring possible relations with the adjusted version of predicted OJA series ($SSM_t = \sum_j \hat{y}_{jt}$) and with its unadjusted raw version ($\text{OJA}_t = \sum_j y_{jt}$) over time.

Specifically, we forecast the LFS series both using only its past by fitting its seasonal autoregressive integrated moving average (SARIMA) and exponential smoothing (Ets) representations and in a bivariate model exploiting possible relationships with unadjusted and adjusted OJA series.

We assessed the order of integration/stationarity of the three series using the augmented Dickey–Fuller (ADF) unit root test (Said and Dickey 1984). To evaluate the existence of a long-term relationship (cointegration), we use the Engle and Granger (1987) two-step strategy, normalizing the first step equation on LFS_t.

Moreover, the best SARIMA and Ets representations were fitted using the `auto.arima` and `ETS` algorithms (Hyndman and Khandakar 2008; Hyndman and Athanasopoulos 2017), based on the minimization of the corrected Akaike information criterion (AICc). If it is the case, when two series are I(1) and cointegrated, we estimate a vector error correction model (VECM, Enders 2010) providing forecasts of LFS that will be compared with two univariate forecasters by comparing forecast accuracy measures, such as root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentual error (MAPE) on the forecast horizon (four quarters ahead).

4 | OJA Data

4.1 | WOLLYBI

Italy has a long history of analyzing data from online job advertisements. An Italian project focused on collecting online job vacancies in Italy from job portals that advertise job advertisements and include newspaper websites, job boards, and employment agencies. This project 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 labor market information system with the aim of collecting, cleaning (accounting for duplicates and errors), and classifying online job vacancies posted on the major Italian websites. In WOLLYBI, OJA were classified according to the standard ISCO-ESCO classification and other dimensions such as sector (Nace), skill requested, education level (ISCED), and region (Nuts 2). The methodological tasks of the knowledge discovery in databases (KDD) approach (Fayyad et al. 1996) for extracting useful and

reliable knowledge from raw data from web sources, including selection of sources, scraping, preprocessing (data cleaning and data deduplication), text mining, and classification, in standard European taxonomy were explained elsewhere (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 (one of the EU's decentralized agencies that supports the European Commission in the field of vocational education and training), 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. Currently, Cedefop collects online job advertisements, as well as experimental data on skills' demand, using the Skills Online Vacancy Analysis Tool for Europe within the web platform Skills-OVATE (link).

The data used in the present paper refers to the Italian web platform WOLLYBI during the period 2013-Q2 to 2018-Q2. The available data are a structured dataset where each observation represents the sum of OJA associated with the following labor market variables:

1. occupation: ISCO-08 (ESCO) classification version (ESCO3, third digit, 130 occupations)
2. territorial unit: Nuts2 (20 Italian regions)
3. aggregate sector of economic activity: Nace Rev.2 (12 sectors)
4. time of the job advertisement (21 quarters, from 2013-Q2 to 2018-Q2)

In its first release, the Italian platform WOLLYBI provided an aggregated classification of economic activity, reducing the 21 NACE categories to 12 broader sectors.

As an illustration, with such data, we can create profiles by cross-classifying these covariates at the most disaggregated level (ESCO3 × Nuts2 × Nace × Quarter), resulting in a total of 1,203,930 profiles. Analyzing data at ESCO2 aggregation instead, we have 379,260 profiles.

4.2 | Skills-OVATE

The aforementioned Skills-OVATE dataset (link), published by Cedefop (a decentralized EU agency that supports the European Commission in vocational education and training), represents a valuable source extending OJA to a broader European context.

Since 2018, Cedefop has been 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). Publicly available OJA are provided at an adequate level of granularity, including the occupational category (ESCO1 and ESCO2) and the region (Nuts2) to which the advertisements pertain. 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 the standard European taxonomy have been explained elsewhere (Cedefop 2019). Skills-OVATE essentially collects the same information provided by WOLLIBY, but unfortunately, the publicly available OJA data from Skills-OVATE are not broken down by all variables of interest but only by Quarters × Nuts2 × ESCO2, and typically, only five quarters of data are provided for each data wave.

In this end, the main advantage of using the WOLLYBI dataset lies in its provision of OJA microdata, detailed across all ESCO3 × Nuts2 × Nace × Quarter levels (covering a period from 2013 to 2018), allowing a more nuanced and disaggregated analysis for representative analyses. In contrast, the public section of Skills-OVATE restricts analysis to a more aggregated level and limits temporal assessment due to the availability of only a few quarters.

Moreover, until 2023, the Skills-OVATE repository also reported 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 the first day the OJA is posted online.

This is important information because it can resolve one of the main problems of OJA, which is to transform OJA flows into stock. This requires assigning each OJA's first posting date to a specific reference quarter-year. Doing so ensures that these 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, 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, a suggested approach in the literature is to assign each OJA to the month (and the relative quarter) in which it remained available/open for the longest duration in the 2-month period.

4.3 | Webographics

Modeling an outcome in a sample selection framework requires for model identification a set of additional covariates, called instruments that enter in the selection equation (probability to observe a nonmissing OJA in a profile of the sample frame) but not the outcome equation (counts of OJA in the profiles).

Literature on OJA suggests achieving the so-called webographics or attitudinal variables, as they capture differences between "online" and "offline" outcome profiles across different territorial units, economic sectors, or population groups. Among these, internet penetration rates (and trends) are particularly noteworthy (Schonlau and Couper 2017; de Pedraza and Serrano 2014). Using data from the Regional Statistics Database (RSD, Eurostat 2023), specifically from

the "Regional Digital Economy and Society" and "Regional Science and Technology Statistics" sections, we selected a set from available webographics (summarized in Table A1 in Appendix 1).

Webographics refer to variables both for citizens (such as internet usage, online home banking, and use of social networks) and for sectors of economic activity (such as the degree of digitalization, measured as a percentage of workers employed in high-tech and knowledge-intensive occupations relative to total employment in a sector).

These series, disaggregated at the NUTS2 level, span different years. However, webographics are only useful when there is enough variability among areas, such as the percentage of people using social networks that varies from 22.2% (NUTS2 region FRY3-Guyane) to 92.8% (DK04-Midtjylland) or the proportion of people using internet home banking ranges from 13.5% (BG42-Yuzhen tsentralen and RO22-Sud Est) to 98.6% (NO07-Nord Norge).

This variability can be useful to find strong instruments, unlike other more generic webographics, such as the percentage of households with internet access at home, which shows negligible variation (ranging from 77.8% in FRY3-Guyane and 82.5% in BG31-Severozapaden to many EU Nuts2 regions with 100%).

4.4 | Empirical Strategy for OJA Modeling and Bias Assessment

Both individual Italian WOLLYBI data (2013-Q2 to 2018-Q2) and the Skills-OVATE dataset (selecting OJA for Spain, Italy, Germany, and France for the quarters of 2022) were manipulated to generate sample frames rising by cross-tabulation of Nuts2 × Nace × ESCO2 × Quarter.

Before fitting the models, we assess the proper distribution of the outcome variable (OJA_j), exploring different options such as Gaussian, lognormal, Poisson, negative binomial, and gamma (Q-Q plots). In both SSM equations and in the OJA equation of multilevel, we specify, as covariates, the sector of activity (Nace), the region (Nuts2), and the best aggregation levels for ESCO (ESCO1 and ESCO2) and time. Regarding time dimension, the large number of available quarters allows to manipulated them differently in each equation (as a continuous variable, or as a 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. Furthermore, the SSM selection equation allows the inclusion of webographics as possible instruments.

Multilevel modeling was estimated using Equation (2), using random intercepts ($\beta_{0j} = \beta_0 + u_{0j}$), where profiles' error terms were normally distributed (u_{0j} distributed as $N(0, \sigma^2)$) and hypothesized to be independent from the observations' error terms e_{jt} (whose distribution depends on the results of the chosen family, based on the Q-Q plot).

The empirical estimation and significance of ρ and the LR test will determine the most suitable methodological approach, such as whether a univariate model (e.g., ML) or a joint model (e.g., SSM) is more appropriate. The main outputs of SSM and ML models are the predictions of OJA counts $E(y_{jt} | \mathbf{x}_{2j})$, denoted by $\hat{y}_{jt}$ that were compared with respect to benchmark data.

We calculated the percentage bias (Bias%) of two models' predictions at time t, aggregated over J profiles ($\text{Pred\_OJA}_t = \sum_j \hat{y}_{jt}$) and the bias of raw OJA (OJA_t), comparing the distributions of these counts with the benchmark data across all J profiles. Specifically, the percentual bias (Bias%) was calculated as $100 \times [(\text{Pred\_OJA}_t - \text{LFS\_New\_Hires}_t)/\text{LFS\_New\_Hires}_t]$, where LFS_New_Hires_t refers to counts of new hires (LFS microdata) at time t, aggregated over all J profiles. Percentage bias was also calculated for marginal distributions, across the categories of auxiliary variables (Quarters, ESCO2, ESCO1, Nuts2, and Nace).

In both applications the LFS New hires as benchmark dataset was obtained from quarterly LFS microdata, as follows: in each dataset (country and quarters of interest for two analyses), 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" (ESCO = 00) and those employed in the sector "Activities of Extraterritorial Organizations and Bodies" (Nace = U). In the LFS microdata, as regional dimension (Nuts2), we considered the region of work (because OJA refers to the region where the job is vacant).

By summing these counts and weighting them by LFS population weights (COEFFQ), we obtained quarterly figures for LFS new hires to use as benchmark.

Regarding data to forecast, LFS recent starters were downloaded from the Eurostat website. This source focuses on new job starters for people aged 15–74 years in Italy from 2013-Q2 to 2018-Q2 (and four quarters ahead for assessing the forecasts) and in the four EU countries (Spain, Italy, Germany, and France) for the quarters of 2022.

5 | Results

Overall, the analyzed Italian 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 (Figure 1), whereas we identified 4,315,814 LFS new hires.

The difference in these figures may seem problematic. However, this presents an opportunity because we have two official sources to compare model predictions. Furthermore, we can assess the bias of our estimates using the benchmark that we consider the most reliable—namely, in our opinion, the LFS new hires. Using this source as a benchmark, these numbers indicate a significant overrepresentation of OJA compared to actual new jobs (as shown in Figure 1), which continuously grew over time.

Table A2 presents univariate marginal distributions of observed OJA by Nuts2, Quarters, ESCO1, and Nace.

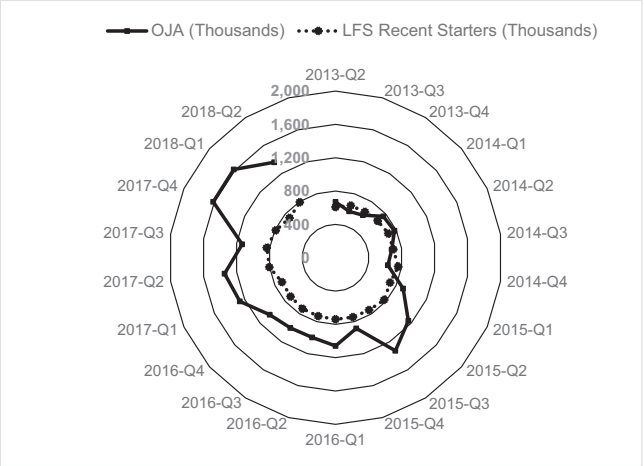

FIGURE 1 | Counts (in thousands) of OJA and LFS recent starters (2013-Q2 to 2018-Q2) in Italy.

TABLE 1 | Missing and nonmissing profiles by ESCO aggregation of 21,403,160 OJA.

	ESCO1	ESCO2	ESCO3
Profiles J	2400	10,320	31,200
%Missing profiles (J)	15.9%	25.4%	42.6%
Observations O ($=J\times$Quarters)	50,400	216,720	655,200
Observations O with OJA >0	30,343	81,934	145,557
%Missing observations	39.8%	62.2%	77.8%

Crossing the levels of Nace (12 aggregated levels), Nuts2 (20 regions), and different aggregations of ESCO (9 levels for ESCO1, 43 levels for ESCO2, and 130 levels for ESCO3) we found a different number (#) of levels for the profiles J or observations O (where $\#O=\#J\times$Quarters), leading those we classed in the sample frame, with a different number of missing (OJA count $=0$) and nonmissing profiles (OJA count >0), as presented in the Table 1.

5.1 | Models' Results (Italian Data)

Studying OJA distributions, we try different theoretical models for their distributions, such as normal, lognormal, gamma, Poisson, and negative binomial. The Q-Q plots suggest transforming OJA counts with natural logarithm (ln_OJA) and assuming that it is Gaussian-distributed (see Figure A1). Thus, all our analyses (ML and SSM) used ln_OJA as the outcome variable, whereas the SSM selection equation was modeled as a Probit model.

Tables A3, A4, and A5 present the estimates of the ML and SSM (the selection and outcome equation, respectively).

As expected, we found a strong positive correlation coefficient between the two equations ($\widehat{\rho}=0.98$***) also confirmed by the LR test (LR chi-square $=5063.4$, $p<0.00001$), implying a strong nonrandom selection, or substantial evidence against the MAR hypothesis, and the two equations cannot be modeled separately. This means that latent factors responsible for OJA sample selection (OJA with positive counts) are the same that drive the number of OJA vacancies in nonempty profiles.

Regarding the selection equation, note the strong significance of instruments, a result that not only legitimates the proper identification of two equations but also has relevance for interpretation.

Namely, OJA are likely to appear online in regions where the population has an advanced "digital" skill (social network participation and home banking). In the same vein, this also occurs for more digitalized regional labor markets in terms of workers' skill (and thus firms) overall considering all sectors (R&D_WORKERS) and especially in the administrative sector (HI_TECH_Empl_N). Interestingly, regions with more digitalized public administration (and agriculture) are less likely to post vacancies online. This may reflect the fact that OJA for the public sector (ministry and institutional web portals) does not enter our data that generally refer to job ads for the private sector.

Table A7 deeply illustrates how the probability of inclusion of the SSM estimated selection equation varies over profiles' characteristics, taking univariate margins or K-way combinations of main covariates.

For example, the minimum mean probabilities of observing a positive count of OJA are as follows: 0.04 for the Molise region, 0.20 for the "Agriculture, forestry, and fishery" sector (Nace), 0.12 for "Chief executives, senior officials, and legislators" (ESCO2 $=11$), and 0.35 in 2013-Q3. The maximum mean probabilities of observing a positive count of OJA are as follows: 0.49 for the Lombardy region, 0.78 for the "Industry and manufacturing" sector, and 0.52 for "Business and administration associate professionals" (ESCO2 $=33$) in the quarter 2018-Q2. Using bivariate counts, the probability of inclusion varies from 0.02 for "Chief executives, senior officials, and legislators" (ESCO2 $=11$) in the Molise region to 0.97 for "Business and administration Associate Professionals" in the Lombardy region.

Using trivariate interactions, the profiles with the lowest mean probability are "Market-oriented skilled forestry/fishery/hunting" (ESCO2 $=63$) or ESCO2 $=11$ in the agriculture sector in Molise (0.02), whereas the highest probability of inclusion (>0.999) refer to the following profiles in quarters 2016-Q2 or 2016-Q3: (1) "Business service agents" in "Manufacturing and industry" or (2) "Real estate; Professional, scientific, and technical; and Administrative and support services" in Lombardy or Veneto region.

Regarding fit statistics, the R-square of 0.671 and RMSE of 1.22 on logged variables demonstrate a quite satisfactory goodness of fit for the SSM outcome (OJA) equation.

Despite the impressive fit statistics of the multilevel equation (R-square $=0.896$, RMSE $=0.43$ on logged scale), the strong selection mechanism found indicates that the two equations cannot be estimated separately. Consequently, this approach

leads to biased and inconsistent parameter estimates and OJA predictions.

5.2 | Bias Assessment

After estimating (exponentiated) predictions of OJA within the profiles using SSM and ML once aggregated over profiles, we compare the total counts of predicted OJA over time (along with the raw counts of OJA) against benchmark data. This comparison is presented for models estimated using data generated by both ESCO 1 and ESCO 2 aggregation levels (see Table 1), which are useful for defining the profiles.

Considering data generated by the first aggregation plan (ESCO1 × Nuts2 × Nace × Quarter), Figure 2 illustrates the evolution of counts (in thousands) of OJA, LFS recent starters, as well as predicted OJA by the SSM and ML. Although multilevel models theoretically should not be shown due to their bias and inconsistency, the plot is included to highlight interesting differences among these models.

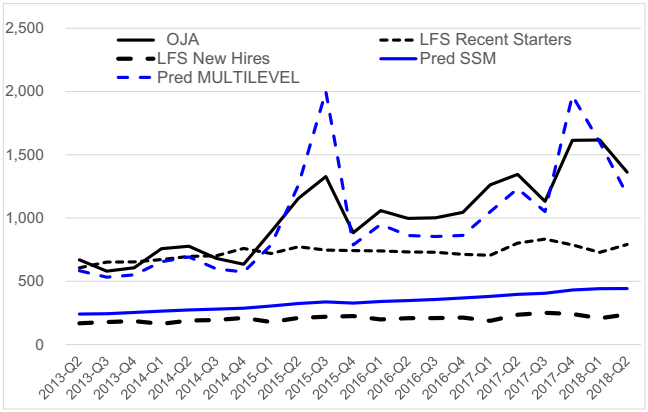

FIGURE 2 | Evolution of LFS recent job starters, LFS new hires, OJA counts, and predicted OJA by SSM and multilevel (in thousands). Profiles created by crossing ESCO1 × Nuts2 × Nace × Quarters.

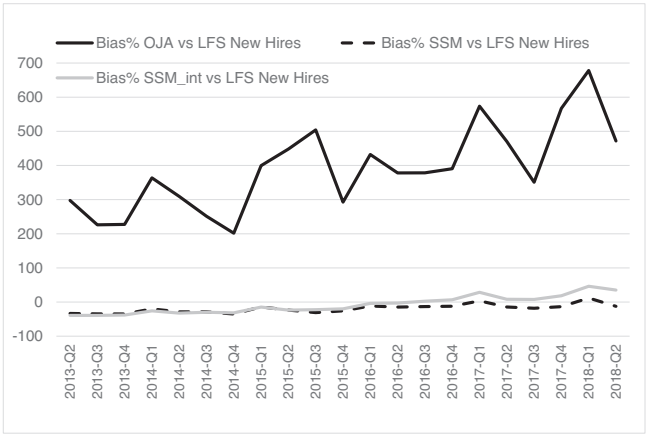

FIGURE 3 | Evolution of Bias% among LFS new hires and OJA, predicted OJA by SSM estimated with interactions (SSM_int) and without (SSM). Profiles created by crossing ESCO2 × Nuts2 × Nace × Quarters.

The multilevel model appears to replicate OJA trajectories over time, amplifying their peaks in the predictions due to the specification of full random terms in the model. In contrast, SSM presents a more smoothed trajectory that aligns more closely with the benchmark (LFS, particularly LFS new hires). Figure 2 demonstrates that the SSM effectively reduces the percentual bias (−23.2%) of OJA relative to the LFS recent starters benchmark (+40.1%) over the 21 quarters, performing significantly better than multilevel (+36.7%).

Specifically, by accounting for both observed and unobserved factors influencing nonrandom selectivity (and the amount of OJA), the predicted SSM counts significantly reduced the raw OJA counts, particularly for profiles more prone to high volumes of OJA.

Regarding data with the second aggregation plan (ESCO2 × Nuts2 × Nace × Quarter), Figure 3 directly shows the percentage bias among predictions and the best benchmark data derived from LFS new hires.

Figure 3 summarizes the very limited bias for both SSM versions with covariate interactions in the OJA equation (−6.5%) and without interactions (−19.1%) over the overall period.

Particularly, summing over quarters and profiles, the total amount of predicted OJA (4,033,907) closely matches the corresponding stock of new hires (4,315,814), unlike the large amount of raw OJA (21,403,160). The multilevel approach is not shown because of excessive bias (Bias% = 384.7%) and the biasedness of its estimates.

These figures confirm that OJA coverage can vary widely over main labor market characteristics and especially over time (Figure 2). Differences in growth emerge when comparing OJA data with stable official sources, demonstrating that SSM represents a promising strategy for addressing and correcting issues of representativeness. Because LFS microdata provides a benchmark for all combinations of Quarters × Nuts2 × Nace × ESCO2, we can control the bias of predictions for each dimension of interest.

Table A6 provides the bias of SSM with covariate interactions by univariate marginals of four auxiliary variables.

SSM adjustments significantly reduce the very large bias observed in raw OJA (+396.9%), particularly for profiles with large stock of OJA, such as for ESCO1, professionals (Bias% OJA = 908.1, Bias% SSM = 122.0), technicians and associate professionals (Bias% OJA = 1277.8, Bias% SSM = 51.4), and clerical support (Bias% OJA = 590.5, Bias% SSM = −5.4). The same error reduction is mainly evident for sectors more prone to appear online, such as ICT (%bias OJA = 2520.3, %bias SSM = 194.5), industry and manufacturing (Bias% OJA = 848.5, Bias% SSM = 58.1), or wholesale and retail (Bias% OJA = 512.7, Bias% SSM = 25.9).

5.2.1 | Cedefop's Skills-OVATE Analysis of 2022 in Four EU Countries

The above analysis was replicated using more recent data from the Skills-OVATE institutional EU repository for the four most populous EU countries. We used the wave that refers to the years

2021–2022 (2021-Q4 to 2022-Q4). Since the beginning of 2023 (Q1), the portal 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. The full analyses and results are reported in Section A2. Most importantly, similar to the Italian results on bias, the SSM predictions for the four countries are highly accurate, closely aligning with the LFS new hires data. However, an exception is noted for Spain, where there is a notable underrepresentation of OJA compared to the LFS benchmark. This discrepancy may be due to the selection mechanisms of OJA on Spanish portals or some characteristics of this labor market.

5.3 | Forecasting Italian Data

Trends in Figure 2 support, at least graphically, the hypothesis that the LFS series and adjusted OJA—exponentiating the predicted by SSM (SSM_t)—may represent the same underlying evolution, useful for forecasting LFS recent starters (LFS_t).

Ignoring the seasonal nature of the series, both the ADF test and the best Arima model found that first differencing the three series LFS_t, OJA_t, and SSM_t results in stationary series, leading to I(1) representation. Instead, when considering seasonal series, results demonstrate a seasonal unit root and a significant drift: More interestingly, both LFS and SSM_t were found with the same best SARIMA representation, with very similar AR1 parameters and drift (Table 2). Based on fit statistics, such a SARIMA model outperforms the best exponential smoothing model (Ets) for the LFS series in the training data (this estimated Ets has multiplicative error, additive trend, and no seasonality, leading to a classical Holt's linear method with multiplicative errors).

Given that LFS_t and SSM_t are I(1) series, they were used to assess possible cointegration, and if it will be the case, adding seasonal dummies in the VECM equations to exploit seasonality. Results show that two series were cointegrated: the estimated cointegrating relationship ($LFS_t = 469038.0\ (51120.4) + 0.8787(0.1714)SSM_t$, with standard errors in parentheses) shows the SSM_t coefficient is strongly significant ($p < 0.0001$, adjusted R-squared $= 0.558$). The residual of the cointegrating relationship was found to be a white noise series, with no significant residual autocorrelation (Ljung–Box test $= 15.742$, $p = 0.7325$).

This implies the existence of a structural long-run relationship, where a one-unit increase in adjusted vacancies results in a less than proportional increase in employment, specifically by a factor of 0.88. The estimated VECM, taking short-term adjustments of only one lag and seasonal dummies into account, is shown in Table 3.

The signs (and magnitude of ECT coefficient for VECM stability) of the coefficients were found to be as expected. Hence, the overall finding is that there is both a strong long-run (contemporaneous between levels) and a moderate short-term (between lags of differences) relation among the series. Furthermore, the numbers employed react to the vacancies but not vice versa.

This implies Granger causation from movement in vacancies to movements in numbers employed. This demonstrates that when demand for labor is strong, the *level* of vacancies will generally rise, which tends to lead to higher employment (*level*), although less proportionally.

Table 4 presents the results of the three forecast models (ETS, SARIMA, and VECM) along with their forecast accuracy, whereas Figure 4 illustrates the forecasts for four quarters ahead.

The VECM emerges as the best-performing model (with an average percentage error of 2.51% for LFS recent starters and a smaller error in the last two quarters), reducing the error—measured by MAPE—by 165% compared to ETS and by 200% compared to SARIMA.

These results were largely expected, given the structural long-run relationship among the LFS series and the updated OJA series.

TABLE 3 | VECM model with error correction term (ECT) and seasonal dummies (Q1, Q2, and Q3).

Covariates	LFS equation	SSM equation
ECT	−0.605(0.278)*	0.066(0.064)
LFS ($t-1$)	0.241(0.260)	0.069(0.060)
SSM ($t-1$)	0.656(1.130)	0.145(0.259)
Q1	−10790.1(15223.4)	8482.4(3545.5)*
Q2	43372.2(18409.9)*	9156.7(4287.6).
Q3	11848.5(21271.2)	570.6(4954.05)

Note: Significant levels: "*" 0.05, "." < 0.1.

TABLE 2 | Seasonal and nonseasonal Arima representation: LFS, OJA, and adjusted OJA (SSM).

	(S)ARIMA model	(S)AR1 coeff. (SE)	Drift	BIC
LFS[a] seasonal	(1,0,0)(0,1,0)[4] with drift	0.590 (0.239)	7648.4 (2315.3)	417.39
LFS[a] no seasonal	(0,1,0)			483.39
OJA seasonal	(0,0,0)(1,1,0)[4] with drift	−0.496 (0.198)	45860.9 (8605.8)	472.04
OJA no seasonal	(0,1,0)			548.94
SSM seasonal	(1,0,0)(0,1,0)[4] with drift	0.572 (0.186)	7647.7 (1049.7)	362.54
SSM no seasonal	(0,1,0)			412.75

[a]Best Ets representation: ETS(M,A,N), BIC = 520.79.

TABLE 4 | LFS, univariate and VECM forecasts, and prediction accuracy metrics.

	Forecast horizon				Forecast accuracy		
	2018-Q3	2018-Q4	2019-Q1	2019-Q2	RMSE	MAE	MAPE
LFS recent starters	745,700	707,800	768,300	805,300			
SARIMA	809,889	811,521	793,209	837,078	64,242.7	56,149.2	7.613
Ets (M,A,N)	792,654	801,745	810,836	819,927	57,125.9	49,515.2	6.731
VECM	774,677	736,747	779,932	810,418	21,442.2	18,668.3	2.531

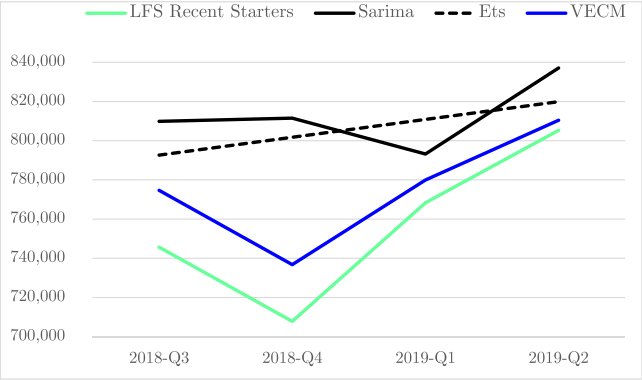

FIGURE 4 | LFS recent starters and forecasting by univariate (SARIMA and Ets) and bivariate VECM (using LFS recent starters and SSM series) models.

6 | Discussion

Given that the observed patterns in OJA data may reflect shifts in the representativeness of different occupational groups over time rather than actual labor market dynamics, this has underscored the need for a suitable statistical approach to correct raw OJA for meaningful near real-time employment forecasts.

Particularly, the revised OJA series, after adjusting for sample selectivity and nonrepresentativeness in new employment dynamics, offers a more accurate depiction of labor demand while also serving as a useful tool for forecasting labor supply data.

The core of the methodological proposal is the creation of the sample frame S_J, derived from the cross-tabulation of the levels of auxiliary variables. These levels correspond to those existing in the population, not just in the individual sample D_I. This approach ensures the availability of two groups within the sample frame: those with missing OJA and those with nonzero OJA (selected profiles). These data augmentation technique is crucial to define the binary outcome for the selection equation, which can be easily modeled because the covariates defining the profiles are fully observed for both groups of observations by construction.

Moreover, reorganizing individual data in such a manner allows (jointly) modeling the missing data mechanism and evaluating possible MAR and MNAR situations, otherwise not possible when working with individual data, because the truncated sample would not provide information on the portion of unsampled observations, as typically occurs in the ML approach. 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). Continuous covariates can be categorized into classes with large granularity or treated as factor variables in the cross-tabulation.

In our specific labor market context, working with OJA sample data that are too much aggregated (e.g., with ESCO1 in combination with other dimensions such as region, sector, and time) may lack profiles with missing OJA, making it impossible to perform a joint SSM (see the upper part of Table 1). Moreover, using ESCO1 may result in missing important information about OJA. In contrast, working with ESCO2 (which includes 40 categories after excluding the three levels related to the armed forces from the total 43 categories) is a better choice. This approach not only provides valuable insights into the occupational dimension of OJA at an appropriate level of aggregation but also reduces the risk of the lack of missing profiles. If this were the case, working with ESCO3 would generate a very large number of profiles, largely increasing the number of missing profiles (see Table 1).

Another important consideration is that the SSM in a GJRM framework may adapt to various outcome distribution scenarios, typically beneficial for count variables, such as Poisson, negative binomial, gamma, or Tweedy (for more extreme situations) distributions. In this end, apart from GJRM, GLLAMMs (Rabe-Hesketh et al. 2004) may be a useful alternative.

6.1 | Potential Benefits of Using Different LFS Sources

From a broader perspective, if we agree that adjusted OJA should mimic and target LFS new hires (or at least consider new hires as an upper bound for OJA), different FLS sources may be used as reference sample. Possible population frames with available auxiliary variables are summarized in Table A9.

The choice of the most appropriate LFS data as a benchmark for bias assessment—and potentially for poststratification—warrants further attention. Because LFS microdata are released with a delay of nearly 2 years, they are unsuitable for near-real-time OJA data analysis and cannot effectively support real-time labor demand analyses. Currently, quarterly LFS data on recent job starters remain a uniquely valuable resource for adjusting OJA data in near-real time. This adjustment is essential

for labor market analyses that model supply factors, since, as illustrated in the paper introduction, they require important demand-side factors, such as demand stocks by occupational groups over time.

However, although LFS recent job starters are disaggregated at the Quarter*x*Nace level, the sectorial disaggregation often contains numerous missing data series, which vary by EU country.

We recommend that LFS data on recent job starters be disseminated at a different level of aggregation, such as Quarter × ESCO1, as this more closely aligns with the key information that OJA are intended to capture (Garasto et al. 2021; Cammeraat and Squicciarini 2021; Turrell et al. 2019; Cedefop 2023, 2024).

This preference is grounded in several practical, methodological, and empirical reasons.

First, occupations are not always concentrated within specific industries, so aggregating occupations by sector may obscure important details. Recalibrated OJA should ideally reflect longitudinal occupational trajectories.

Second, the structure and information content of OJA make classification by occupation more precise than classification by sector of economic activity. Typically, OJA reports the sector of the enterprise posting the advertisement (Boselli et al. 2017; Cammeraat and Squicciarini 2021) rather than the sector of the job being advertised. This discrepancy raises concerns about potential measurement error biases, which should be examined alongside nonresponse biases (Zhang 2005).

Third, another motivation arises from our empirical analysis showing significant differences between LFS recent starters and LFS new hires. This suggests that LFS job starters may include irrelevant new starters, complicating the accurate adjustment of OJA and bias correction for key dimensions such as occupation and NUTS. For example, Figure 5 compares the stocks of OJA from the Skills-OVATE portal for various ESCO1 groups with our selected LFS new hires for the last quarter of 2022.

At first glance, Figure 5 reveals that, contrary to expectations, online job postings do not necessarily indicate higher job demand for individuals with higher skill levels and more advanced educational backgrounds (such as professionals) in Germany. This is especially evident in France, where the volume of OJA for Elementary occupations exceeds that for technicians–associate professionals and service–sales occupations.

Apart from these considerations, Figure 5 highlights significant misalignment between the observed OJA and reference data by occupation (a dimension not available in the LFS recent starters).

Thus, it is evident that the current version of LFS recent job starters, without further disaggregation, cannot be adequately used for either bias assessment or effective poststratification strategies. Consequently, to accurately adjust OJA stocks through poststratification, it is essential to have benchmark data with ESCO1-level disaggregation.

Fourth, as previously discussed, variance reduction in the presence of MNAR mechanisms can be achieved during the poststratification step by selecting proper auxiliary variables. It is generally advisable to choose auxiliary variables for the population frame that are either the most significant or the most correlated with the outcome, especially when a model-based strategy is absent (poststratification of the observed, rather than predicted, outcome variable) or when using less sophisticated models that often result in poor fit (Gelman 2007; Warshaw and Rodden 2012). Conversely, if the missingness model is a poor fit or if irrelevant covariates are used for defining profiles, generating accurate predictions from selected samples can be challenging, even under the MAR assumption (Buttice and Highton 2013; Lax and Phillips 2009).

To this end, occupation generally exhibits higher statistical significance compared to sector. This finding has been confirmed

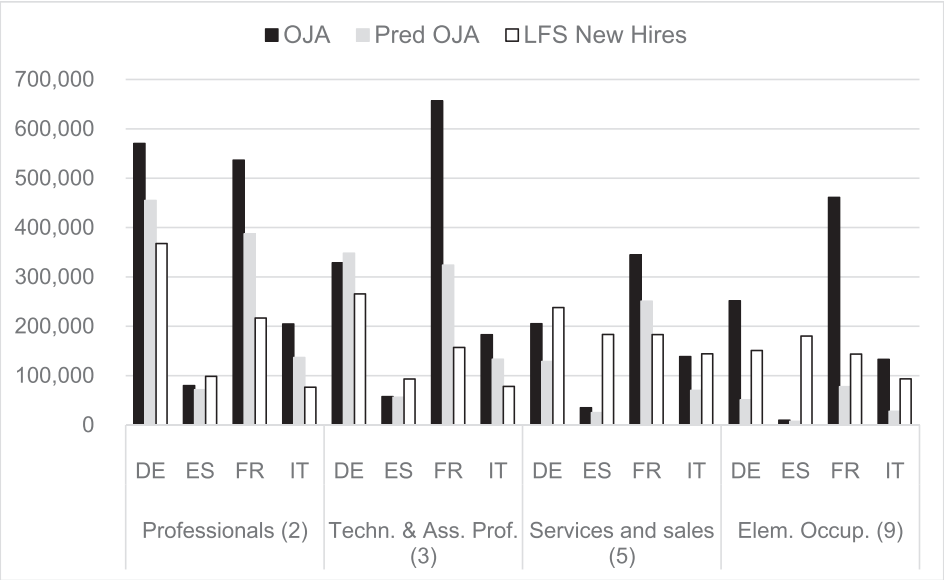

FIGURE 5 | Distributions of total OJA, predicted OJA (by SSM), and total LFS new hires in 2022-Q4, by country and some ESCO1 group (ESCO1 code in parentheses).

by our study, as well as other empirical applications (Cammeraat and Squicciarini 2021; Turrell et al. 2019).

6.2 | Limitations

The paper acknowledges several limitations.

First, a limitation of the SSM version used in this paper is that it does not incorporate pseudo-random effects, thereby failing to exploit the hierarchical data structure (time dependencies within profiles).

Unlikely the adopted specification (where longitudinal observations were not correlated within the profiles), GJRM may exploit the data hierarchy (time nested in profiles) by specifying the same effects that drive the hierarchy in a multilevel model (adding in both Equations 1 and 2 a sum of J profiles, as dummy variables, multiplied by their coefficients), not as random effects (thus avoiding new random errors in both equations), but as penalized fixed effects whose coefficients were estimated under some penalization constraints. Because generally, the profiles are very numerous, the use of penalized coefficients is justified because we suspect that some coefficients may be weakly or not identified or underrepresented in the data. Notably, 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, 2011), exactly like the way they are estimated in multilevel models (known as empirical Bayes or BLUP), but specified as random terms.

Despite such limitations, we believe that addressing selectivity is the primary concern for joint models aimed at adequate labor market analysis, as opposed to univariate approaches such as ML that, even if managed nested data, produce biased and inconsistent estimates.

Secondly, the absence of the most recent data for WOLLIBY and Cedefop's Skills-OVATE has hindered more up-to-date analyses.

Third, the limited number of available quarters for Cedefop's Skills-OVATE data and the unavailability of public data because 2023-Q1 have prevented a forecast analysis for EU countries. In this perspective, past publicly available information from Skills-OVATE is only broken down by quarters, NUTS2, and ESCO2. This lack of detailed data poses a significant issue for making precise, real-time adjustments to OJA for labor demand analysis. It hinders the accurate quarterly refinement of OJA stocks. The absence of a sectoral dimension limits our ability to investigate recent demand trends, such as the progress of digitalization in both society and the public sector. For example, analyzing the OJA data for ICT professionals and associate professionals within public administration would provide insights into regional variations across the EU.

7 | Conclusion

The paper introduces a promising approach to correct OJA from nonrandom selectivity and nonrepresentativeness to be used to forecast occupational dynamics. This method allows for the construction of more consistent measures of vacancy distributions that mimic official employment data, also providing more accurate forecasts.

Model-based inference and poststratification are standard techniques for addressing unequal probabilities of selection and nonresponse, provided the nonresponse mechanism is known or partially accounted for by covariates. However, these approaches assume the MAR mechanism.

In contrast, our paper adopts a more conservative strategy based on econometric literature on SSMs, which can address potential MNAR mechanisms. This method accounts for unobservable or unmeasured variables that may influence both nonresponse and the outcome variable.

Concerning the forecast results, previous econometric estimates suggest that the addition of lagged values of vacancy stocks to simple autoregressive models or multivariate time series models improves their explanatory power in relation to employment growth (European Central Bank 2002; Australian Bureau of Statistics 2003; Lovaglio et al. 2020). Specifically, results indicate that quarterly growth in job vacancies, *rather than exhibiting contemporaneous movement*, leads employment growth by between one and three-quarters.

The present approach demonstrates that adjusting OJA for nonrandom selectivity reveals both a long-run (contemporaneous) and a short-term relationship between the series, which can be leveraged to forecast LFS recent starters.

These results suggest that a deeper analysis of the comovement of two series could contribute to better forecasts of employment data in a near real-time fashion.

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 nonresponse bias for estimator of target variable totals, also in the presence of MNAR) and, although not fully developed in the present paper, a poststratification strategy (leading always to variance reduction in case of a high association between auxiliary variables of the population frame and the outcome).

In contrast, the strength of the multilevel method lies in the potential for further poststratification. However, we argue that poststratifying biased estimates may not be the correct strategy. In this end, there is a notable parallelism between MRP and SSM regarding the need for auxiliary variables. Both methodologies require these variables but for different purposes: MRP needs them to improve the fit of the outcome equation, whereas SSM enriches the possible population profiles in order to better model nonresponse.

However, we question whether the true goal of strategies addressing nonrandom selectivity is merely a horse race to enhance the predictive power of a model, given that the observed outcomes stem from a truncated and nonrepresentative sample.

Similarly, problematic situations involving incomplete matching in ML (e.g., having more profiles in the population frame than

in the sample frame) can be addressed as opportunities within an SSM framework.

In conclusion, the key methodological challenge for future labor market analyses lies in integrating web-based data with official statistics to produce timely and cost-effective measurements. Achieving accurate inferences will require both reliable near real-time benchmark sources and robust statistical methods capable of addressing the inherent biases and gaps in online job advertisement data.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available because of privacy or ethical restrictions.

References

Amoah, B. 2000. "Help-Wanted Index. Perspectives on Labour and Income." Statistics Canada, 75-001-XPE, 14–18.

Australian Bureau of Statistics. 2003. Do Job Vacancies Provide a Leading Indicator of Employment Growth? Australian Labour Market Statistics. Catalogue no. 6105.0, April 2003.

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.

Boselli, R., M. Cesarini, F. Mercorio, and M. Mezzanzanica. 2014. "Planning Meets Data Cleansing." *Proceedings of the International Conference on Automated Planning and Scheduling* 24: 439–443.

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.

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.

Caines, C., F. Hoffmann, and G. Kambourov. 2017. "Complex-Task Biased Technological Change and the Labor Market." *Review of Economic Dynamics* 25: 298–319.

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

Card, D., and A. B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84, no. 4: 772–793.

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. http://data.europa.eu/doi/10.2801/438491.

Cedefop. 2024. Untangling Labour Shortages in Europe: Unmet Skill Demand or Bad Jobs? http://data.europa.eu/doi/10.2801/023297.

Chetty, R., A. Guren, D. Manoli, and A. Weber. 2011. "Are Micro and Macro Labour Supply Elasticities Consistent? A Review of Evidence on the Intensive and Extensive Margins." *American Economic Review* 101, no. 3: 471–475.

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." Working Paper 154, Amsterdam Institute for Advanced Labour Studies, Amsterdam.

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. J. 2017. "The Growing Importance of Social Skills in the Labor Market." *Quarterly Journal of Economics* 132, no. 4: 1593–1640.

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.

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

Enders, W. 2010. *Applied Econometric Time Series*. 3rd ed. John Wiley & Sons.

Engle, R. F., and C. W. J. Granger. 1987. "Co-Integration and Error Correction: Representation, Estimation, and Testing." *Econometrica* 55: 251–276.

European Central Bank. 2002. "Monthly Bulletin." European Central Bank, Frankfurt.

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.

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.

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.

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, 8 March 2021.

Gelman, A. 2007. "Struggles With Survey Weighting and Regression Modeling." *Statistical Science* 22, no. 2: 153–164. https://doi.org/10.1214/088342306000000691.

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 [May 2017].

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.

Haggar-Guénette, C. 1989. Jobs Ads: A Leading Indicator? Perspectives on Labour and Income. Statistics Canada, 1, 75-001-XPE, 1–12.

Hamermesh, D. S. 1993. "Labor Demand: What Do We Know? What Don't We Know?" *Journal of Labor Economics* 11, no. 1: 1–39.

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.

Hyndman, R. J., and G. Athanasopoulos. 2017. *Forecasting: Principles and Practice*. 2nd ed. OTexts.

Hyndman, R. J., and Y. Khandakar. 2008. "Automatic Time Series Forecasting: The Forecast Package for R." *Journal of Statistical Software* 27: 1–22.

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.

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. 2012. "The Democratic Deficit in the States." *American Journal of Political Science* 56, no. 1: 148–166.

Lee, L. F. 1983. "Generalized Econometric Models With Selectivity." *Econometrica* 51, no. 2: 507–512.

Lichter, A., A. Peichl, and S. Siegloch. 2015. "The Own-Wage Elasticity of Labour Demand: A Meta-Regression Analysis." *European Economic Review* 80: 94–119.

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.

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.

Lovaglio, P. G. 2025. "Cross-Learning With Panel Data Modeling for Stacking and Forecast Time Series Employment in Europe." *Journal of Forecasting* 44, no. 2: 753–780.

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* 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.

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

Mandrone, E., M. Tancioni, and S. Laj. 2010. "Vacancies as Employment Predictor: Dynamic Properties of the Series'." In *Job Vacancies in the Italian Labour Market. The New ISFOL "Help Wanted" Time Series, ISFOL "TEMI E RICERCHE"*, edited by E. Mandrone. Rubettino.

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.

Moffitt, R. 1984. "The Estimation of a Joint Wage-Hours Labour Supply Model." *Journal of Labor Economics* 2, no. 4: 550–566.

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.

Neumark, D., and W. L. Wascher. 2007. "Minimum Wages and Employment." *Foundations and Trends in Microeconomics* 3, no. 1–2: 1–182.

Newey, W. K., J. L. Powell, and J. R. Walker. 1990. "Semiparametric Estimation of Selection Models: Some Empirical Results." *American Economic Review* 80, no. 2: 324–328.

OECD. 2023. *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market*. OECD Publishing. https://doi.org/10.1787/08785bba-en.

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.

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

Radice, R., G. Marra, and M. Wojtys. 2016. "Copula Regression Spline Models for Binary Outcomes." *Statistical Computing* 26, no. 5: 981–995.

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

Rosen, S., and F. Welch. 1971. "Labour Supply and Income Redistribution." *Review of Economics and Statistics* 53, no. 3: 272–283.

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

Ruth, F. J., J. G. Schouten, R. F. Wekker. 2006. "Statistics Netherlands Business Cycle Tracer. Methodological Aspects; Concept, Cycle Computation and Indicator Selection." Statistics Netherlands Discussion Paper, 5, 1–64.

Said, S. E., and D. A. Dickey. 1984. "Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order." *Biometrika* 71: 599–607.

Salvatori, A. 2018. "The Anatomy of Job Polarisation in the UK." *Journal for Labour Market Research* 52: 8. https://doi.org/10.1186/s12651-018-0242-z.

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.

Sklar, A. 1973. "Random Variables, Joint Distributions, and Copulas." *Kybernetica* 9: 449–460.

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

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., and D. Zimmer. 2007. "Copula Modeling: An Introduction for Practitioners." *Foundations and Trends(R) 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." National Bureau of Economic Research.

Turrell, A., J. Thurgood, D. Copple, J. Djumalieva, 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.

Valletta, R. 2005. "Help-Wanted Advertising and Job Vacancies." *Federal Reserve Bank of San Francisco Economic Letter* 2: 1–3.

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.

Vella, F. 1998. "Estimating Models With Sample Selection Bias: A Survey." *Journal of Human Resources* 33, no. 1: 127–169.

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. 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: 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.

Zagorsky, J. L. 1998. "Job Vacancies in the United States: 1923 to 1994." *Review of Economics and Statistics* 80: 338–345.

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.

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.

Appendix

Section A1: WOLLIBY (Italian Data) Results

TABLE A1 | Webographics from the Regional Statistics Database (link).

GoodsServices_WEB	Percentage of individuals who online purchase in the last 3 months (378 Nuts2 annual series 2006–2023)
REGULAR_WEB	Percentage of individuals regularly using the internet (214 Nuts2, annual series 2012–2023)
HOME_WEB	Percentage of households with internet access at home (378 Nuts2, annual series 2006–23)
HOME_BANK	Percentage 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, and sharing (378 Nuts2 annual series 2006–2023)
AuhorityInteract_WEB	Percentage of individuals who used the internet for interaction with public authorities (359 Nuts2, annual series 2008–21)
NEVER_WEB	Percentage of those who have never used a computer (350 Nuts2 annual series 2006–2017)
MOBILE_WEB	Percentage of individuals who accessed the internet away from home or work (357 Nuts2, annual series 2012–19)
R&D_WORKERS	Percentage of FTE workers in R&D on labor force (business enterprise sector) (627 Nuts2, annual series 1980–2022)
HI_TECH_Empl_j	Percentage of employment in technology and knowledge-intensive occupations of total employment for Sector *j* (505 Nuts2, annual series 2013–2023)

TABLE A2 | %OJA distribution by Nuts2, Nace, Quarters (Q), and ESCO1, overall period 2013-Q2 to 2018-Q2.

Nuts2	%	Nace/ESCO	%	Q	%
Lombardia	32.8	Industry and manufacturing ($C+B+D+E$)	33.0	2013-Q2	3.1
Veneto	14.2	Real estate + Professional, scientific and technical + Administrative and support service ($L+M+N$)	21.7	2013-Q3	2.7
Emilia Romagna	13.7	Wholesale retail trade and repair of motor vehicles (G)	15.6	2013-Q4	2.8
Piemonte	9.2	Information and communication (J)	9.0	2014-Q1	3.5
Lazio	6.9	Transportation and storage (H)	5.2	2014-Q2	3.6
Toscana	6.7	Accommodation and food service (I)	4.5	2014-Q3	3.2
Friuli V.G.	2.6	Other service activities + Households as employers ($S+T$)	4.1	2014-Q4	3.0
Marche	2.3	Education + Health/social work + Arts/entertainment ($P+Q+R$)	2.9	2015-Q1	4.2
Campania	2.0	Financial and insurance (K)	2.6	2015-Q2	5.4
Liguria	1.7	Construction (F)	0.6	2015-Q3	6.2
Puglia	1.5	Agriculture, forestry, and fishery (A)	0.5	2015-Q4	4.1
Abruzzo	1.4	Public administration and excluded defense (O)	0.3	2016-Q1	4.9
Sicilia	1.2			2016-Q2	4.7
Trentino	1.2	**ESCO1**	%	2016-Q3	4.7
Umbria	1.0	1 Managers/directors	2.7	2016-Q4	4.9
Sardegna	0.6	2 Professionals	18.1	2017-Q1	5.9
Calabria	0.4	3 Technicians and associate professionals	26.3	2017-Q2	6.3
Basilicata	0.2	4 Clerical support	14.6	2017-Q3	5.3
Molise	0.2	5 Service and sales	10.6	2017-Q4	7.5
Val d'Aosta	0.2	6 Skilled agriculture/forestry/fishery	0.1	2018-Q1	7.6
		7 Craft and related trade	12.6	2018-Q2	6.4
		8 Plant and machine	9.0		
		9 Elementary occupations	6.2		

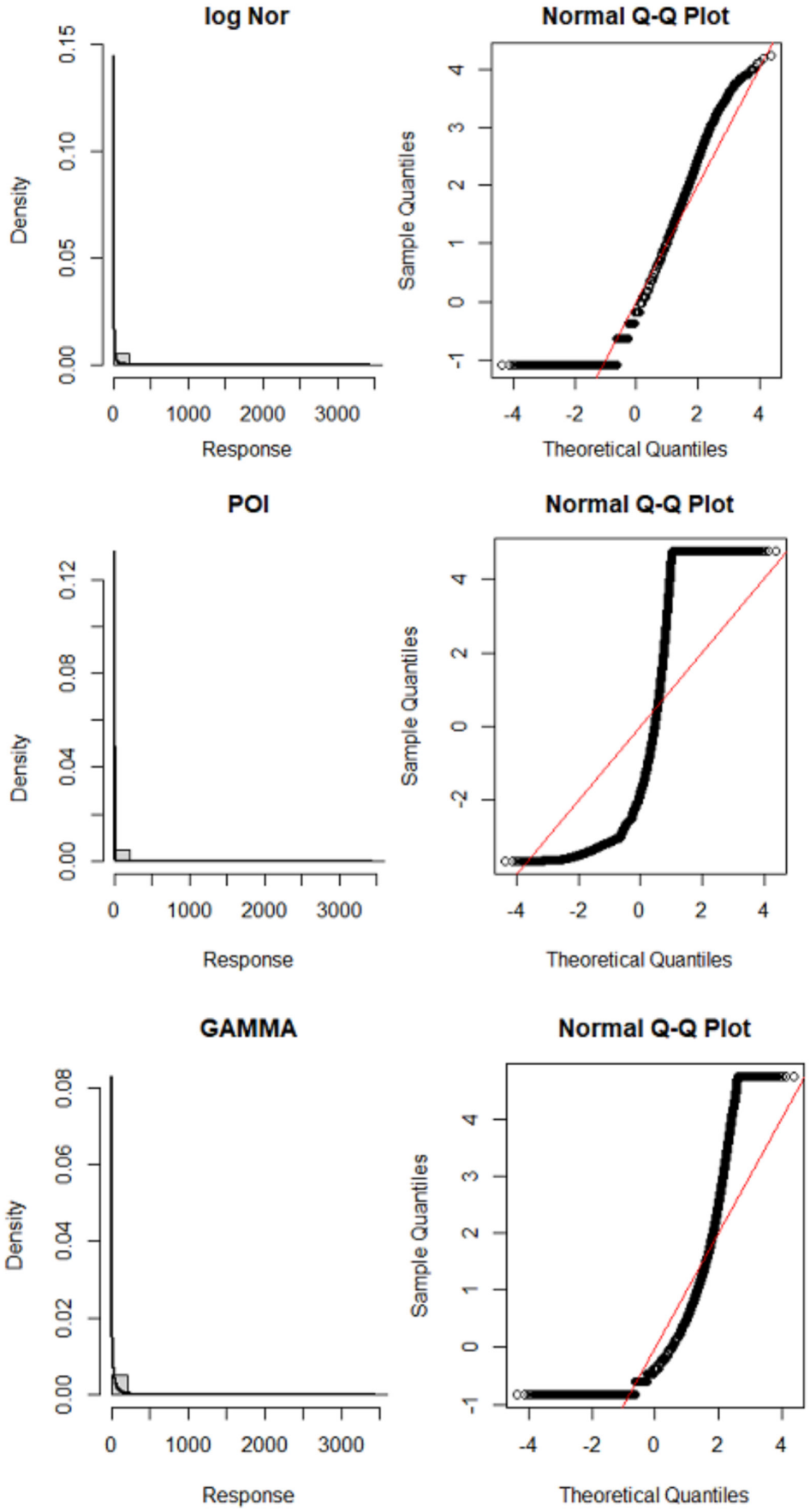

FIGURE A1 | Distributions and Q-Q plot of OJA distributions (Italy, 2013-Q2 to 2018-Q2) under the hypothesis of lognormal (log nor), Poisson (POI), and gamma distribution (Italy, Quarters 2013–2018). *Note:* The Q-Q plot of the negative binomial is almost identical to that of the Poisson.

TABLE A3 | Multilevel results of the ln_OJA equation, intercept as random effects (Italy, 2013-Q2 to 2018-Q2).

OJA equation	Estimate	St. error	*t* value	Pr > |*t*|
Intercept	5.16	0.113	45.8	< 0.0001
ESCO1 1. Managers	−0.41	0.078	−5.2	< 0.0001
ESCO1 2. Professionals	0.94	0.076	12.3	< 0.0001
ESCO1 3. Technicians and associate professionals	1.46	0.076	19.3	< 0.0001
ESCO1 4. Clerical support workers	0.91	0.076	12.0	< 0.0001
ESCO1 5. Service and sales workers	0.48	0.076	6.2	< 0.0001
ESCO1 6. Skilled agricultural, forestry, and fishery workers	−2.07	0.094	−22.2	< 0.0001
ESCO1 7. Craft and related trade workers	0.24	0.076	3.2	0.0016
ESCO1 8. Plant and machine operators and assemblers	−0.14	0.077	−1.9	0.0641
ESCO1 9. Elementary occupations	0.00	.	.	.
ITC1 Piemonte	0.09	0.113	0.8	0.4221
ITC2 Valle d'Aosta	−2.56	0.121	−21.2	< 0.0001
ITC3 Liguria	−1.09	0.114	−9.6	< 0.0001
ITC4 Lombardia	1.21	0.112	10.8	< 0.0001
ITF1 Abruzzo	−1.29	0.114	−11.3	< 0.0001
ITF2 Molise	−2.56	0.121	−21.2	< 0.0001
ITF3 Campania	−0.95	0.113	−8.4	< 0.0001
ITF4 Puglia	−1.18	0.114	−10.4	< 0.0001
ITF5 Basilicata	−2.36	0.119	−19.9	< 0.0001
ITF6 Calabria	−2.05	0.118	−17.4	< 0.0001
ITG1 Sicilia	−1.36	0.114	−12.0	< 0.0001
ITG2 Sardegna	−1.73	0.115	−15.0	< 0.0001
ITH12 Trentino Alto-Adige	−1.38	0.115	−12.0	< 0.0001
ITH3Veneto	0.40	0.113	3.6	0.0003
ITH4 Friuli Venezia Giulia	−0.96	0.114	−8.5	< 0.0001
ITH5 Emilia Romagna	0.40	0.113	3.5	0.0004
ITI1 Toscana	−0.15	0.113	−1.3	0.1816
ITI2 Umbria	−1.53	0.115	−13.3	< 0.0001
ITI3 Marche	−1.02	0.113	−9.0	< 0.0001
ITI4 Lazio	0.00	.	.	.
2013-Q2	−0.71	0.023	−30.9	< 0.0001
2013-Q3	−0.80	0.023	−34.6	< 0.0001
2013-Q4	−0.76	0.023	−33.1	< 0.0001
2014-Q1	−0.59	0.023	−26.0	< 0.0001
2014-Q2	−0.54	0.023	−23.6	< 0.0001
2014-Q3	−0.68	0.023	−30.2	< 0.0001
2014-Q4	−0.72	0.023	−31.6	< 0.0001
2015-Q1	−0.41	0.022	−18.4	< 0.0001
2015-Q2	0.06	0.022	3.0	0.0029
2015-Q3	0.65	0.021	30.3	< 0.0001

(Continues)

TABLE A3 | (Continued)

OJA equation	Estimate	St. error	*t* value	Pr > |*t*|
2015-Q4	−0.41	0.022	−18.2	< 0.0001
2016-Q1	−0.22	0.022	−10.0	< 0.0001
2016-Q2	−0.32	0.022	−14.2	< 0.0001
2016-Q3	−0.33	0.022	−14.6	< 0.0001
2016-Q4	−0.32	0.022	−14.2	< 0.0001
2017-Q1	−0.12	0.022	−5.4	< 0.0001
2017-Q2	0.04	0.022	1.8	0.0651
2017-Q3	−0.12	0.022	−5.4	< 0.0001
2017-Q4	0.51	0.021	23.8	< 0.0001
2018-Q1	0.30	0.022	14.0	< 0.0001
2018-Q2	0.00	.	.	.
Nace1 Agriculture	−1.49	0.092	−16.2	< 0.0001
Nace2 Industry and manufacturing	1.55	0.087	17.8	< 0.0001
Nace3 Construction	−1.44	0.092	−15.7	< 0.0001
Nace4 Wholesale and retail	1.21	0.087	13.9	< 0.0001
Nace5 Transportation and storage	−0.10	0.088	−1.1	0.2548
Nace6 Accommodation and food	−0.15	0.089	−1.6	0.1004
Nace7 ITC	−0.22	0.090	−2.4	0.0156
Nace8 Financial and insurance activities	−0.53	0.089	−6.0	< 0.0001
Nace9 Real estate + Professional/scientific/technical + Administrative activities	1.10	0.087	12.6	< 0.0001
Nace10 Public administration	−1.63	0.092	−17.7	< 0.0001
Nace11 Education + Health/social + Arts/entertainment	−0.36	0.090	−4.1	< 0.0001
Nace12 Other services + Activities of households as employment	0.00	.	.	.

Note: Variance of random intercept = 0.6441, $p < 2e\text{-}16^{***}$. Estimates 0.00 indicate the reference level of each categorical variable.

TABLE A4 | Results of SSM for the probit selection Equation (1).

Selection equation	Label	Estimate	St. error	*z* value	Sign.
Intercept		−69.05	0.162	−426.0	< 0.0001
Nace	Agriculture	−1.66	0.030	−56.3	< 0.0001
	Industry and manufacturing	1.55	0.029	53.6	< 0.0001
	Construction	−1.46	0.029	−50.4	< 0.0001
	Wholesale and retail	1.22	0.027	44.7	< 0.0001
	Transportation and storage	−0.03	0.026	−1.0	0.3230
	Accommodation and food	−0.07	0.027	−2.5	0.0113
	ITC	−0.19	0.027	−7.3	< 0.0001
	Financial and insurance activities	−0.62	0.027	−23.0	< 0.0001
	Real estate + Professional/scientific/technical + Administrative activities	1.09	0.027	40.3	< 0.0001
	Public administration	−1.96	0.030	−65.1	< 0.0001
	Education + Health/social + Arts/entertainment	−0.33	0.027	−12.4	< 0.0001
	Other services + Activities of households as employment	0.00	.	.	.
Nuts2	Piemonte	0.29	0.089	3.3	0.0010
	Valle d'Aosta	−2.01	0.064	−31.5	< 0.0001
	Liguria	−0.70	0.059	−11.8	< 0.0001
	Lombardia	0.89	0.088	10.1	< 0.0001
	Abruzzo	−0.18	0.081	−2.2	0.0311
	Molise	−1.23	0.098	−12.5	< 0.0001
	Campania	0.23	0.070	3.2	0.0013
	Puglia	0.62	0.111	5.6	< 0.0001
	Basilicata	−0.77	0.111	−7.0	< 0.0001
	Calabria	0.12	0.141	0.9	0.3860
	Sicilia	0.41	0.102	4.0	< 0.0001
	Sardegna	−0.59	0.073	−8.1	< 0.0001
	Trentino Alto-Adige	−0.42	0.091	−4.6	< 0.0001
	Veneto	0.80	0.091	8.9	< 0.0001
	Friuli Venezia Giulia	−0.43	0.063	−6.9	< 0.0001
	Emilia Romagna	0.66	0.094	7.0	< 0.0001
	Toscana	0.26	0.069	3.8	0.0002
	Umbria	−0.79	0.075	−10.6	< 0.0001
	Marche	−0.52	0.081	−6.4	< 0.0001
	Lazio	0.00	.	.	.
ESCO1	Managers	−0.38	0.024	−16.1	< 0.0001
	Professionals	0.95	0.023	41.1	< 0.0001
	Technicians and associate professionals	1.47	0.024	60.9	< 0.0001
	Clerical support	0.95	0.023	41.2	< 0.0001
	Service and sales	0.48	0.023	20.9	< 0.0001
	Skilled agriculture/forestry/fishery	−2.84	0.033	−85.3	< 0.0001

(Continues)

TABLE A4 | (Continued)

Selection equation	Label	Estimate	St. error	*z* value	Sign.
	Craft and related trades	0.23	0.023	9.9	< 0.0001
	Plant and machine operators/assemblers	−0.24	0.023	−10.5	< 0.0001
	Elementary occupations	0.00	.	.	.
Year		0.03	0.000	.	.
Instruments					
HOME_BANK	%People using home banking	0.02	0.002	9.5	< 0.0001
SOCIAL_WEB	%People using social networks	0.02	0.002	9.4	< 0.0001
R&D_WORKERS	%Workers in R&D on labor force	0.21	0.067	3.1	0.0022
HI_TECH_Empl_AB[a]	AB (agriculture and mining)	−0.16	0.013	−12.2	< 0.0001
HI_TECH_Empl_N[a]	N (administrative activities)	0.15	0.013	11.6	< 0.0001
HI_TECH_Empl_O[a]	O (public administration)	−0.02	0.010	−2.3	0.0204

[a]Percentage of employment in ICT + knowledge-intensive occupations of total employment for this sector. Estimates 0.00 indicate the reference level of each categorical variable.

TABLE A5 | Results of SSM for the ln_OJA Equation (2).

OJA equation	Label	Estimate	St. error	*t* value	Sign.
Intercept		4.98	0.045	109.8	< 0.0001
Nace	Agriculture	−2.18	0.034	−64.7	< 0.0001
	Industry and manufacturing	1.89	0.030	63.8	< 0.0001
	Construction	−1.91	0.033	−58.2	< 0.0001
	Wholesale and retail	1.51	0.030	51.1	< 0.0001
	Transportation and storage	−0.05	0.030	−1.6	0.1172
	Accommodation and food	−0.12	0.030	−4.0	< 0.0001
	ITC	−0.25	0.031	−8.1	< 0.0001
	Financial and insurance activities	−0.77	0.031	−24.9	< 0.0001
	Real estate + Professional/scientific/technical + Administrative activities	1.34	0.030	45.1	< 0.0001
	Public administration	−2.48	0.035	−71.7	< 0.0001
	Education + Health/social + Arts/entertainment	−0.43	0.031	−13.9	< 0.0001
	Other services + Activities of households as employment	0.00	.	.	.
Nuts2	Piemonte	0.21	0.038	5.6	< 0.0001
	Valle d'Aosta	−3.19	0.044	−73.2	< 0.0001
	Liguria	−1.14	0.039	−29.4	< 0.0001
	Lombardia	1.44	0.038	38.5	< 0.0001
	Abruzzo	−1.49	0.039	−37.8	< 0.0001
	Molise	−3.41	0.045	−75.7	< 0.0001
	Campania	−1.04	0.039	−26.9	< 0.0001
	Puglia	−1.36	0.039	−34.7	< 0.0001
	Basilicata	−3.09	0.043	−71.4	< 0.0001
	Calabria	−2.68	0.042	−64.2	< 0.0001
	Sicilia	−1.64	0.040	−41.4	< 0.0001
	Sardegna	−2.13	0.040	−52.6	< 0.0001
	Trentino Alto-Adige	−1.58	0.039	−40.0	< 0.0001
	Veneto	0.58	0.038	15.4	< 0.0001
	Friuli Venezia Giulia	−0.94	0.039	−24.4	< 0.0001
	Emilia Romagna	0.62	0.038	16.4	< 0.0001
	Toscana	−0.04	0.038	−1.0	0.3414
	Umbria	−1.79	0.040	−44.9	< 0.0001
	Marche	−1.14	0.039	−29.4	< 0.0001
	Lazio	0.00	.	.	.
ESCO1	Managers	−0.50	0.027	−18.3	< 0.0001
	Professionals	1.21	0.026	46.2	< 0.0001
	Technicians and associate professionals	1.84	0.026	71.0	< 0.0001
	Clerical support	1.19	0.026	45.8	< 0.0001
	Service and sales	0.61	0.026	23.0	< 0.0001
	Skilled agriculture/forestry/fishery	−3.55	0.039	−91.7	< 0.0001

(Continues)

TABLE A5 | (Continued)

OJA equation	Label	Estimate	St. error	*t* value	Sign.
	Craft and related trades	0.30	0.027	11.1	<0.0001
	Plant and machine operators/assemblers	−0.29	0.027	−10.9	<0.0001
	Elementary occupations	0.00	.	.	.
Quarters[a]	2013-Q2	−0.84	0.032	−26.7	<0.0001
	2013-Q3	−0.85	0.032	−26.7	<0.0001
	… ..				
	2016-Q3	−0.25	0.030	−8.5	<0.0001
	… ..				
	2017-Q3	−0.07	0.029	−2.5	0.0117
	2017-Q4	0.04	0.028	1.4	0.1597
	2018-Q1	0.19	0.028	6.8	<0.0001
	2018-Q2	0.00	.	.	.

[a]Empty quarters' estimates exhibited the same trend as nonempty quarters in the preceding periods. Estimates 0.00 indicate the reference level of each categorical variable.

TABLE A6 | Bias% of SSM prediction (model with interactions) with respect to LFS new hires, by covariates' levels. Italy, 2013-Q2 to 2018-Q2.

Var	Label	LFS new hires	Pred OJA	Bias%
Quarter	2013-Q2	168,421	102,587	−39.1
	2013-Q3	177,869	108,482	−39.0
	2013-Q4	185,368	114,734	−38.1
	2014-Q1	163,548	121,363	−25.8
	2014-Q2	189,806	128,394	−32.4
	2014-Q3	193,946	135,852	−30.0
	2014-Q4	210,421	143,763	−31.7
	2015-Q1	178,493	152,157	−14.8
	2015-Q2	210,981	161,063	−23.7
	2015-Q3	219,779	170,515	−22.4
	2015-Q4	225,014	180,546	−19.8
	2016-Q1	198,990	191,194	−3.9
	2016-Q2	208,570	202,498	−2.9
	2016-Q3	209,412	214,499	2.4
	2016-Q4	212,986	227,243	6.7
	2017-Q1	187,422	240,776	28.5
	2017-Q2	235,816	255,149	8.2
	2017-Q3	251,055	270,416	7.7
	2017-Q4	241,903	286,635	18.5
	2018-Q1	207,791	303,867	46.2
	2018-Q2	238,224	322,176	35.2
ESCO1	1. Managers	19,243	657,204	3315.4
	2. Professionals	384,282	853,000	122.0
	3. Technicians and associate professionals	406,985	616,100	51.4
	4. Clerical support	452,560	427,920	−5.4
	5. Service and sales	1,196,651	372,163	−68.9
	6. Skilled agriculture/forestry/fishery	45,817	243,177	430.8
	7. Craft and related trades	585,615	353,726	−39.6
	8. Plant and machine operators/assemblers	318,910	185,560	−41.8
	9. Elementary occupations	905,751	325,057	−64.1
Nuts2	Piemonte	289,490	350,952	21.2
	Valle d'Aosta	11,630	18,975	63.2
	Liguria	100,495	107,684	7.2
	Lombardia	699,544	1,156,120	65.3
	Abruzzo	92,892	82,345	−11.4
	Molise	15,749	18,853	19.7
	Campania	331,151	119,569	−63.9
	Puglia	299,450	88,009	−70.6
	Basilicata	40,253	22,218	−44.8
	Calabria	137,693	30,683	−77.7

(Continues)

TABLE A6 | (Continued)

Var	Label	LFS new hires	Pred OJA	Bias%
	Sicilia	291,217	69,391	−76.2
	Sardegna	165,603	47,345	−71.4
	Trentino Alto-Adige	104,601	73,172	−30.1
	Veneto	408,787	487,796	19.3
	Friuli Venezia Giulia	91,290	119,478	30.9
	Emilia Romagna	367,101	491,683	33.9
	Toscana	285,075	274,909	−3.6
	Umbria	70,974	62,627	−11.8
	Marche	114,858	108,177	−5.8
	Lazio	397,961	303,922	−23.6
Nace	Agriculture	281,281	38,489	−86.3
	Industry and manufacturing	744,625	1,177,425	58.1
	Construction	328,649	47,894	−85.4
	Wholesale and retail	544,968	685,817	25.9
	Transporting and storage	203,719	177,026	−13.1
	Accommodation and food	651,534	158,880	−75.6
	ITC	73,514	216,500	194.5
	Financial and insurance activities	35,791	126,823	254.4
	Real estate + Professional/scientific/ technical + Administrative activities	396,981	951,965	139.8
	Public administration	55,303	32,946	−40.4
	Education + Health/social + Arts/entertainment	739,724	190,201	−74.3
	Other services + Activities of households as employment	259,726	229,940	−11.5

Section A2: Cedefop's Skills-OVATE Analysis in Four EU Countries

Following the suggested approach in the literature, we assigned each OJA to the month (and the relative quarter) in which it remained available/open for the longest duration in the 2-month period. Columns 3 and 4 in Table A8 present the results of the recent recalibration, which adjusts the OJA data for the quarters of 2022. This adjustment allows for the exclusion of over 6.5 million (6,608,000) OJA entries that pertain to irrelevant periods (2021-Q4 or 2023-Q1). Moreover, examining the marginal figures in these columns confirms that the SSM predictions are more accurately aligned with LFS new hires compared to LFS recent starters.

Although the latter group has a comparable amount of OJA, the similarity is coincidental rather than indicative of a meaningful comparison, due to the inability to further disaggregate this total stock beyond the time dimension.

TABLE A7 | Mean probability of OJA selection (mean *p*) by SSM without interactions, along with the lowest and highest values, and the coefficient of variation (CV) by covariate.

Vars.	Mean *p*	CV
Nuts2 (region)		
Molise	0.04	212.3
Valle d'Aosta	0.05	212.7
Basilicata	0.06	204.9
… … … … … .	… … … … … .	… …
Veneto	0.38	92.6
Emilia Romagna	0.38	91.2
Lombardia	0.49	77.6
Nace (sector)		
Agriculture, forestry, and fishery	0.05	222.9
Construction	0.08	191.4
… … … … … .	… … … … … .	… … …
Wholesale and retail trade and Repair of motor vehicles	0.36	94.3
Real estate; Professional, scientific, and technical; and Administrative and support services	0.39	91.3
Industry and manufacturing	0.44	85.1
Quarter		
2013-Q3	0.16	153.5
2013-Q4	0.17	151.6
2013-Q2	0.17	148.8
… … …	… … …	… … …
2017-Q1	0.22	131.4
2015-Q2	0.29	113.3
2016-Q3	0.35	99.7
ESCO2 (occupation)		
Chief executives, senior officials, and legislators (11)	0.13	418.2
Market-oriented skilled forestry/fishery/hunting (63)	0.13	418.2
… … … … ..	… … … … … .	… …
Business and administration professionals (24)	0.62	54.1
Customer service clerks (42)	0.66	44.3
Business service agents (33)	0.78	33.5

(Continues)

The first column of Table A8 summarizes the results of the four fitted SSM in each country in terms of the (chi-square) LR test on rho's significance and instruments found.

It is noteworthy that, with the exception of Italy, a strong nonrandom selection mechanism was identified, particularly in Germany and France. Webographics proved to be very useful and significant instruments for Italy and France, because they have adequate variability among their regions, unlike Spain and Germany. In this last country, instead, a finer regional disaggregation (Nuts2) and a continuous version of quarters (time) were employed as instruments in the selection equation, whereas Nuts1 and Quarter (specified as class variable) were used in the OJA equation. For Spain, we found only the Quarter as instrument.

Columns 5, 6, and 7 of Table A8 present the total predicted OJA alongside the totals from two benchmark datasets: recent job starters and new hires, segmented by quarters for each country. SSM predictions are generally quite accurate, closely matching the LFS new hires data, except for Spain, where the observed discrepancies warrant further investigation.

TABLE A8 | OJA by grab_date quarter, OJA reassigned to the correct quarter (refined), LFS new hires, LFS recent starters, and predicted OJA by SSM (in thousands), by country and quarter.

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Country[a] (LR test on rho = 0)	**Quarter**	**OJA by grab_date**	**OJA refined**	**LFS new hires**	**LFS recent starters**	**Pred OJA SSM[b]**
DE	2021-Q4	1807				
Chi-square = 925.2***	2022-Q1	1986	1746	1533	1727	1449
Instruments: Nuts2 and Nuts2*Time	2022-Q2	1846	1849	1629	1751	1616
	2022-Q3	2150	2096	1616	1706	1934
	2022-Q4	1723	1982	1497	1667	1775
ES	2021-Q4	235				
Chi-square = 3.04*	2022-Q1	261	232	695	1255	146
Instruments: Quarter	2022-Q2	288	245	875	1241	160
	2022-Q3	262	300	820	1236	176
	2022-Q4	222	258	765	1195	194
FR	2021-Q4	3708				
Chi-square = 19.7***	2022-Q1	2376	2535	999	1315	1423
Instruments: ESCO1*Home_Bank	2022-Q2	2137	2354	1088	1296	1433
	2022-Q3	2026	1813	1047	1221	1162
	2022-Q4	2999	3018	1048	1237	1826
IT	2021-Q4	1024				
Chi-square = 1.35 ($p = 0.25$)	2022-Q1	798	736	599	897	456
Instruments: Hi_TECH_EMPL_K (financial and insurance act)	2022-Q2	881	857	677	840	496
	2022-Q3	872	815	593	773	539
	2022-Q4	823	978	613		586
Total		**28,424**	**21,816**	**16,092**	**20,142**	**15,371**

[a]Below each country, the statistic of the LR test (and significance) for the absence of sample selection and the instruments used in the SSM selection equation (not entering the OJA equation) are provided.
[b]The data used are OJA (refined), once they have been assigned to correct quarters, as explained in the text.

Section A3: LFS Data as Benchmark Dataset

TABLE A9 | Official sources potentially useful for the OJA population frame.

Dataset	Broken down by	Delay (dissemination)	Website
LFS public quarterly data	Country, quarter, sex, **Nace**, and tenure of the last job <3 and ≥ 3 months (**new hires**)	≤ 2 quarters of delay	link
LFS public quarterly data	Country, quarter, and tenure of the last job (<3 months) by sex and age (**recent job starters**)	≤ 2 quarters of delay	link
LFS microdata	Country, quarter, sex, age, Nace, ESCO (three digits), and tenure of the last job	≥ 2 years of delay	link
Eurostat regional statistics	Country, **Nuts2**, year, age, sex, and tenure of the last job (<11 months, 12–13, 24–59, and >60)	>1 year of delay	link