Received 15 October 2023
Revised 22 February 2024
Accepted 18 March 2024

Labour market segmentation and the gender wage gap in Spain

Fernando Núñez Hernández
Department of Industrial Organization and Business Management, University of Seville, Sevilla, Spain
Carlos Usabiaga
Universidad Pablo de Olavide, Sevilla, Spain, and
Pablo Álvarez de Toledo
Department of Industrial Organization and Business Management, University of Seville, Sevilla, Spain

Abstract

Purpose – The purpose of this study is to analyse the gender wage gap (GWG) in Spain adopting a labour market segmentation approach. Once we obtain the different labour segments (or idiosyncratic labour markets), we are able to decompose the GWG into its observed and unobserved heterogeneity components.
Design/methodology/approach – We use the data from the Continuous Sample of Working Lives for the year 2021 (matched employer–employee [EE] data). Contingency tables and clustering techniques are applied to employment data to identify idiosyncratic labour markets where men and/or women of different ages tend to match/associate with different sectors of activity and occupation groups. Once this "heatmap" of labour associations is known, we can analyse its hottest areas (the idiosyncratic labour markets) from the perspective of wage discrimination by gender (Oaxaca-Blinder model).
Findings – In Spain, in general, men are paid more than women, and this is not always justified by their respective attributes. Among our results, the fact stands out that women tend to move to those idiosyncratic markets (biclusters) where the GWG (in favour of men) is smaller.
Research limitations/implications – It has not been possible to obtain remuneration data by job-placement, but an annual EE relationship is used. Future research should attempt to analyse the GWG across the wage distribution in the different idiosyncratic markets.
Practical implications – Our combination of methodologies can be adapted to other economies and variables and provides detailed information on the labour-matching process and gender wage discrimination in segmented labour markets.
Social implications – Our contribution is very important for labour market policies, trying to reduce unfair inequalities.
Originality/value – The study of the GWG from a novel labour segmentation perspective can be interesting for other researchers, institutions and policy makers.

Keywords Oaxaca–blinder decomposition, Gender wage gap, Clustered contingency tables, Labour matching heatmap, Segmented Spanish labour market

Paper type Research paper

We thank the R + D + i Project PROYEXCEL-00724 (Andalusian Board) as well as the PAIDI SEJ 513 Research Group (Andalusian Board) for the funding provided. We also greatly acknowledge the suggestions by the Editor and Reviewers of the journal, as well as the comments by Raquel Llorente, José María Arranz and Raúl Ramos. Funding for open access publishing: Universidad Pablo de Olavide/CBUA.

International Journal of Manpower
Vol. 45 No. 10, 2024
pp. 16-41
Emerald Publishing Limited
0143-7720
DOI 10.1108/IJM-10-2023-0601

1. Introduction

When economists wonder why men receive higher wages than women or try to analyse how this wage gap evolves over time or behaves within the wage distribution itself, they usually resort to the so-called decomposition methods. The driving authors of this methodology are Oaxaca (1973) and Blinder (1973) – both papers focus on the issue of wage discrimination. The basic idea of the Oaxaca–Blinder (OB) decomposition method consists of answering the following two questions: (1) How large is the part of the gender wage gap (GWG) that can be attributed to gender differences in those characteristics that are relevant to wages? This portion (due to the endowments of each one) is called the "explained" component of the gap. (2) How large is the part of the GWG which is due to differences in how those relevant characteristics are rewarded in the labour market for men and women? This second part of the gap (due to the coefficients/returns of each one) is called the "unexplained" component. In the context of the GWG, this second component is often interpreted as "discrimination," at least partially (Jann, 2008) though there is not always discrimination behind it.

Although the Spanish labour market is showing a positive evolution after the COVID-19 pandemic, with 21 million employed people and an unemployment rate of 11.6% in the second quarter of 2023, the truth is that it continues to be a problematic market characterised by elevated long-term unemployment, high youth unemployment, strong segmentation, low regional and occupational mobility and wage inequality between men and women. To address the issue of gender pay inequality, in this paper, we combine the OB decomposition technique with the empirical framework of clustered contingency tables (CTs). We use labour-matching data from a large database of administrative records, the *Continuous Sample of Working Lives* (*Muestra Continua de Vidas Laborales*, MCVL) and structure them into a clustered CT that cross-classifies the information on gender and age of workers and occupation group and activity sector of job placements. The clustered CT allows us to represent the labour market in a segmented way. From this table, we can obtain a heatmap of the labour market that shows how workers, depending on their gender and age, tend to associate with different occupation groups and sectors of activity in the labour market. This heatmap allows us to identify idiosyncratic labour markets of men or women or both where pay inequalities can be very different. In our opinion, the problem of wage inequality cannot be approached as a problem of the labour market as a whole but has to be analysed in a segmented way, looking at each labour market segment. This segmented vision of labour matching in Spain will help to better design measures against the GWG.

There are several theoretical models that can offer support to our empirical analysis, such as the theories of labour market segmentation; the two-sided matching models (Gale and Shapley, 1962; Roth and Sotomayor, 1992), where the occupations and activities chosen by men and women would fundamentally depend on their respective preferences; or the models where companies have sufficient market power to allow themselves to discriminate against certain groups of workers. In this last line, we highlight Becker's (1971) "discrimination of taste" model. Becker (1971) stated that an aversion felt by employers toward persons belonging to certain groups might constitute a source of discrimination and lead to lower wages for discriminated workers. He presented this hypothesis in formal terms by assuming that the gains these employers derive from employing workers include the profit of the firm and some taste parameters. However, such discrimination cannot persist under perfect competition, as employers with no preference will drive employers with discriminatory preferences out of the market by offering all workers equal wages. Hence, the presence of imperfect competition in the labour market is necessary to explain the existence and persistence of discrimination. In this sense, limitations on personal mobility permit firms to exercise monopsonistic power and to pay workers with identical productive abilities differently (Cahuc *et al.*, 2014, Ch. 8). This low-mobility argument has been used to explain

discrimination against women and certain ethnic minorities (see, for example, Gordon and Morton, 1974; Barth and Dale-Olsen, 2009). Our research hypothesis is that the mobility limitations of workers between occupation groups and/or sectors of activity in Spain may be giving companies market power in certain labour market segments and the opportunity to discriminate against workers according to their preferences.

There is literature that uses decomposition methods to analyse the GWG in Spain (and other countries). For instance, Hidalgo (2010) uses quantile regression to simulate counterfactual densities and decompose the Spanish wage inequality evolution (period 1980–2000) into changes due to coefficients, endowments and non-observable worker characteristics (following the methodologies of Machado and Mata, 2005; Autor *et al.*, 2006, 2008). The data used are coming from the *Household Budget Survey* and the MCVL. According to this author, wage inequality follows a counter-cyclical trend from the mid-eighties onwards and changes in both, coefficients and endowments, play an important role in this evolution. Guner *et al.* (2014) observed a GWG in Spain of around 20% in 2010, a figure quite close to its value in 1994. The authors use data from the Spanish *Labour Force Survey* (*Encuesta de Población Activa,* EPA) from 1977 to 2013. Using two decomposition methods (OB and decomposition using quantile regression) and considering the problem of sample selection bias, they observe that the GWG is driven mainly by differences in returns to individual characteristics – women are more qualified than men in observable labour market characteristics but earn less. The same techniques are implemented by Dueñas and Moreno (2018), which analysed the GWG in the Spanish, French and German labour markets in 2015 using microdata from the *EU Statistics on Income and Living Conditions* (EU-SILC, 2016). The results obtained indicate that Spain is the country with the lowest wage gap and the biggest wage discrimination, Germany being the country with the biggest wage gap and the lowest wage discrimination and France in an intermediate position. Thus, in the case of Spain, practically the whole GWG is due to the unexplained part of the OB model (98.29%) – this percentage is lower in the cases of France (72.84%) and Germany (58.13%). For their part, Murillo-Huertas *et al.* (2017) examine regional differences in the GWG in Spain using matched employer–employee (EE) microdata: 2002, 2006 and 2010 waves of the *Survey of Earnings Structure* (*Encuesta de Estructura Salarial*, EES). Their findings suggest that Spain shows a significant regional heterogeneity in the size of the raw gap. Their OB decomposition analysis shows that although the bulk of the GWG in Spanish regions is due to differences in the endowments of productive characteristics between males and females, there is still a substantial part of the gender gap that remains unexplained.

There is also literature for the Spanish economy on wage differentials between groups or collectives other than men and women but where the gender variable plays an important role as a control variable. For example, by adopting a regional perspective, García and Molina (2002) show that being a man reduces the wage gap between the different regions analysed (North, East, South or Centre of Spain) and Madrid. Such a discriminatory effect is higher in the North, East and South than in the Centre of Spain – on interregional wage differentials in Spain, see also Murillo-Huertas *et al.* (2020). In the field of labour insertion, we can highlight the paper of Arrazola *et al.* (2022). These authors show that there are gender differences in the labour insertion process of recent Spanish graduates. These differences are, in general, systematically negative for women and are especially important for the salaries received and the type of contract (part-time/full-time, temporary/permanent contract), although they also depend on the branch of knowledge of the studies – for example, in the engineering branch, the gender gap in the probability of having a relatively high salary (which is unfavourable to female graduates) is explained almost entirely by unobservable institutional or socioeconomic factors, i.e. by the unexplained part of the probability gap. Another wage gap analysed for the Spanish economy is the one that arises when comparing wages in the private sector and the public sector. For instance, this gap is analysed by Couceiro de León and Dolado (2023). These authors find that those unobserved female characteristics which

increase the probability of working in the public sector have a favourable impact on wages, but this public wage premium is only observed for low-educated women. These findings are in part consistent with those of Hospido and Moral-Benito (2016) who find positive selection towards the public sector at the bottom of the wage distribution – in this field, see also the study of Antón and Muñoz de Bustillo (2015).

From the reviewed literature, at least three conclusions can be drawn: (1) It is evident that there is a wage gap unfavourable to women in the Spanish labour market and that a significant part of this gap is due to unobserved factors that affect either the constant of the Mincer wage equation or the return of their explanatory variables. (2) The use of the Spanish MCVL in GWG analysis is not common, and even though these data contain remuneration information at the individual level and a wide range of explanatory variables for that remuneration. (3) The reviewed literature clearly shows that the seminal methodological contributions of Oaxaca (1973) and Blinder (1973) have given way to a broad set of methodological advances that allow decomposition methods to adapt to different data structures and research questions – see the survey by Fortin *et al.* (2011). Among other improvements in the decomposition methodology, we can mention the following: incorporation of standard errors and confidence intervals into the estimates; contributions of simple covariates to the explained and unexplained components of the gap; treatment of dummy variables as explanatory variables; decomposition of discrete choice models; correction of sample selection bias (Heckman, 1979); decomposition of differences in mean outcome differentials (Smith and Welch, 1989; Juhn *et al.*, 1991); combination of decomposition and matching techniques (Nopo, 2008); gap decomposition along the wage distribution using quantile information (Machado and Mata, 2005; Melly, 2005; Firpo *et al.*, 2018); and decomposition for panel data and mixed models (Smith and Welch, 1989; Kim, 2010; Kröger and Hartmann, 2021).

In many of these methodological contributions underlies, in one way or another, the segmentation of the population (or the sample) analysed. Some studies approach segmentation exogenously (according to external classifications), for example, segmenting the sample by earning quantiles (Juhn *et al.*, 1993), or analysing the GWG for different groups in the labour market (by race, region, period of time, public or private job, etc.) leading to an analysis of the gap between the gender gaps of the groups analysed (Smith and Welch, 1989; Juhn *et al.*, 1991). On the other hand, other authors approach segmentation endogenously (i.e. using information from the sample/population to make the segmentation), for example, admitting that there are workers with different probabilities of accessing a job and, therefore, a salary (Heckman, 1979), or looking for groups of men and women which have comparable characteristics, giving rise to the OB analysis of the common support of the distributions of observable characteristics (Nopo, 2008).

To our knowledge, none of the proposed segmentations control for the existence of labour segments where the propensity of young/older men/women to be matched to jobs may differ significantly. Our labour market segments are based not directly on spatial criteria (as, for example, in Manning and Petrongolo, 2017) but on how men and women of different age groups are matched with the different sectors of activity and occupation groups existing in the labour market. Combining CT and clustering methodologies, we can identify labour segments where men (or women) of certain ages show a high propensity to match/associate with certain sectors of activity and occupation groups – on this methodology see the works of Álvarez de Toledo *et al.* (2018), (2020). The fact that young/older men/women show different propensities to match up with certain activity sectors and occupation groups can be due to three reasons: (1) their respective preferences when looking for a job, (2) the preferences of companies when hiring them and (3) their search patterns and those of the companies that hire them (geographical search areas, search channels used, etc.).

Our statistical segmentation procedure is guided by an economic criterion: who matches with whom in the labour market. We want to relate the wage gap of each labour segment (or idiosyncratic market) with the propensity to match/associate men and women of different

ages with the jobs offered in that segment. In principle, it would be expected that in those labour segments where the wage gap is more favourable to young (older) men, the propensity of young (older) women to match is lower and vice versa.

The rest of the paper is structured as follows: Section 2 presents our labour market segmentation methodology based on clustered CTs. Section 3 describes the labour-matching data used (MCVL). Section 4 offers the results of estimating a wage equation and decomposing the GWG adopting a segmented vision of the Spanish labour market. Finally, Section 5 shows the general conclusions of our work.

2. Labour market segmentation methodology

The use of CTs makes sense when there are categorical variables that provide relevant information about a phenomenon under study (Mosteller, 1968; Agresti, 2013). Each cell of the CT shows the frequency of a particular combination of categories of the different categorical variables that are cross-represented in it. From this starting table, correspondence (association) and clustering analyses can be carried out to obtain a new table containing the degree of association between the different categories of the categorical variables; we will call this table the heatmap.

The heatmap proposed in this study allows us to characterise the employment episodes of the MCVL by measuring the degree of association between the different categories of four categorical variables that have an important weight when it comes to segmenting the labour market; namely, gender and age group of the worker and occupation group and activity sector (industry) of the job. To obtain this heatmap, the first step is to cross-classify these four variables in a CT where the rows represent the combinations of the categories of the occupation and industry variables, and the columns represent the combinations of the categories of the gender and age group variables (see Table 1). Our economic hypothesis is that there are certain rows and columns of this CT which tend to be strongly associated (they tend to appear together in the CT) and this should be reflected in the heatmap; in other words, certain young/older men/women tend to match with job vacancies belonging to certain occupational groups and activity sectors. When we refer to strong association, we do not necessarily mean that a particular combination of categories of the four variables considered has a relatively high frequency in the CT, but that this frequency is higher than the one expected if the generation of category combinations were totally random.

By crossing 10 occupation groups and 10 activity sectors, 100 rows are generated in the CT and heatmap. By crossing 2 gender categories with 8 age groups, 16 columns are generated in the CT and heatmap. Therefore, these two tables have a total of 1,600 cells (16 rows × 100 columns). This high number of cells makes it difficult to identify association patterns in the heatmap of the labour market. To overcome this problem, we smooth the

Table 1. Two-dimensional smoothed (unclustered) CT of employment episodes (four categorical variables)

		Y categories (cross-classification of 2 genders and 8 age groups)						
X categories (cross-classification of 10 occupation groups and 10 activity sectors)		1	2	…	*j*	…	*16*	Total
	1	$n_{1,1}$	$n_{1,2}$	…	$n_{1,j}$	…	$n_{1,16}$	n_{1+}
	2	$n_{2,1}$	$n_{2,2}$	…	$n_{2,j}$	…	$n_{2,16}$	n_{2+}
	…			…		…		
	i	$n_{i,1}$	$n_{i,2}$	…	$n_{i,j}$	…	$n_{i,16}$	n_{i+}
	…			…		…		
	100	$N_{100,1}$	$n_{100,2}$	…	$n_{100,j}$	…	$n_{100,16}$	n_{100+}
	Total	n_{+1}	n_{+2}	…	$n+j$	…	n_{+16}	n

Source(s): Authors' own work based on MCVL

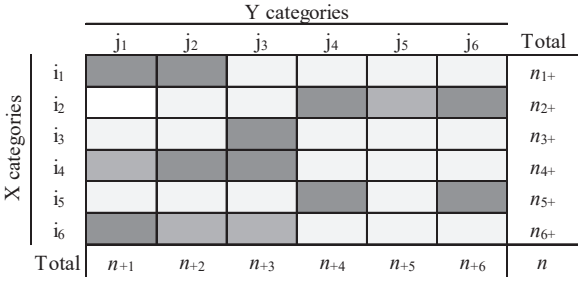
observed CT and apply a clustering procedure to its two sides; in this way, the analysis focuses on homogeneous groups of rows and columns rather than on individual rows and columns. The procedure to obtain the smoothed ordered (clustered) CT and the corresponding heatmap can be seen in Álvarez de Toledo *et al.* (2018, 2020). Here we summarise it in the following steps:

1. The observed CT is smoothed resulting in Table 1. Smoothing techniques in the CT framework provide solutions for estimating cell frequencies (and their probabilities) in the presence of "sparsity." When in a CT the number of cells is too high and/or the finite sample is too small, some cells with positive occurrence probabilities can be zero or have a very small frequency – this phenomenon is known as a zero frequency problem or sparsity problem. In this scenario of sparse high-dimensional CTs, multivariate statistical analyses (as, for example, correspondence analyses, association factors or $\chi 2$ tests of independence) may lose the optimal properties that they show in larger samples.

2. From the smoothed and unclustered CT (Table 1), two auxiliary tables are generated that, respectively, collect the observed and expected probabilities of having an employment episode in each cell – note that these two probabilities are calculated on the already smoothed CT. The observed probability in each cell i,j comes from the quotient $n_{i,j}/n$, while the expected one is obtained as the product of the row marginal probability n_{i+}/n and the column marginal probability n_{+j}/n corresponding to each cell i,j.

3. The quotient of both auxiliary tables allows for obtaining a table of association factors (a_{ij}) between rows and columns [1]: $a_{ij} = \frac{\frac{n_{ij}}{n}}{\frac{n_{i+}}{n}.\frac{n_{+j}}{n}} = n\frac{n_{ij}}{n_{i+}\cdot n_{+j}}$. Factor values higher than one would mean that the association between the corresponding row and column of the table is greater than in a random assignment scenario and vice versa. As an example, Figure 1 represents an association table with an arbitrary order of six rows and six columns. For better visualisation, we have coloured the cells according to the values of the association factor, the higher the factor, the darker the cell.

4. The CT is clustered on the row side and on the column side. The hierarchical (average linkage) clustering methodology is based on a similarity measure between the elements that are clustered (row categories or column categories). We measure the similarity between each pair of rows of the CT (i_A and i_B) as the overlapping or percentage of coincidence of their respective row profiles $\frac{n_{i_Aj}}{n_{i_A+}}$ and $\frac{n_{i_Bj}}{n_{i_B+}}$ with each of the different column categories j: $sim_{i_A\_i_B} = \sum_j min\left(\frac{n_{i_Aj}}{n_{i_A+}}, \frac{n_{i_Bj}}{n_{i_B+}}\right)$. This measure of similarity moves by definition between 0 and 1 and can be calculated in an analogous way on the column side; i.e. two



Source(s): Authors' own work based on MCVL

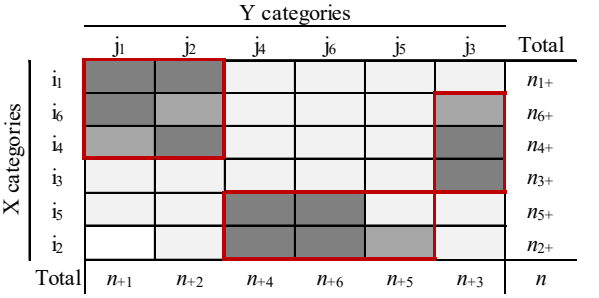
Figure 1. Unclustered association table

column categories of the CT are more similar, the more they resemble the way they are matched with the row categories.

5. The clustering process on both sides of the CT gives rise to separate row and column dendrograms. The dendrogram graphically shows how the row (or column) categories are joined sequentially to give rise to homogeneous groups or clusters of categories. By definition, the base of each dendrogram (the one with the rows and the one with the columns) places the respective clustered categories by proximity. So, these respective bases can be used to order the rows and columns of the association table giving rise to a clustered association table or heatmap (or "gravity" map) that makes it possible to get a panoramic view of how certain groups/clusters of rows tend to be associated with certain groups/clusters of columns and vice versa. Figure 2 shows the association table from Figure 1 after it has been sorted using the information of the clustering process. As can be seen, this figure shows the existence of row clusters and column clusters. For example, rows i_1, i_6 and i_4 form a row cluster because they are similar in the way they are associated with the columns of the table, and columns j_1 and j_2 form a column cluster because they resemble the way they are associated with the table rows.

Our segmentation scheme is not incompatible with the existence of occupational/sectoral mobility in the labour market; in fact, the existence of certain mobility between nearby occupations and/or activities may be favouring the formation of the clusters that we observe in the heatmap (for example, the mobility of older men between the agriculture and construction sectors in certain regions of Spain). It is also true that if disruptive changes in mobility patterns were observed, the heatmap would change its shape, but this type of change occurs slowly because it requires the retraining of workers. In any case, there is evidence of low occupational and sectoral mobility in Spain, which gives our heatmap stability in the short term – see, for example, Anghel *et al.* (2020) and Fernández-Cerezo and Montero (2021) at sectoral level and Caparrós-Ruiz (2016) and Bisello *et al.* (2022) at occupational level.

It can happen that a cluster of rows tends to be associated with a cluster of columns and vice versa. This case is called a bicluster (or labour market segment) and can be explored for idiosyncratic features – three possible biclusters have been marked with red borders in Figure 2. For example, we could look inside a particular bicluster to analyse its structure by gender and age of the worker, region of the workplace, activity sector and occupation group of the job placement, worker earnings, etc. In this study, we will focus on analysing the wage gap by gender.



Source(s): Authors' own work based on MCVL

Figure 2.
Clustered association table or heatmap

The GWG can be analysed by following the OB decomposition. This statistical method is used to analyse the differences in mean outcomes between two groups. As can be seen in Eq. (1), the OB decomposition of the wage differential helps researchers to understand the extent to which differences in observed characteristics ("endowments" or explained factors) and differences in the returns of those characteristics ("coefficients" or unexplained factors, including discrimination) contribute to the overall GWG. For example, we can measure whether men earn more than women because they have more experience in the labour market or because the same level of experience is more highly valued in the case of men. For its part, the "interaction" component shows a simultaneous effect of differences in endowments and coefficients, and it usually presents a small or negligible effect on the explained differential.

$$\overline{W}_m - \overline{W}_f = \underbrace{(\beta_{0m} - \beta_{0f}) + \sum_k (\beta_{km} - \beta_{kf})\overline{X}_{kf}}_{\text{(Coefficients)}} + \underbrace{\sum_k \beta_{kf}\left(\overline{X}_{km} - \overline{X}_{kf}\right)}_{\text{(Endowments)}} + \underbrace{\sum_k (\beta_{km} - \beta_{kf})\left(\overline{X}_{km} - \overline{X}_{kf}\right)}_{\text{(Interactions)}} \quad (1)$$

where m: male; f: female; $\overline{W}$: average wage; k explanatory variables (excluding the intercepts β_{0f} and β_{0m}); $\overline{X}_k$: average values of the explanatory variables; and β_k: Mincer estimated coefficients.

Note that the Mincer equation, estimated, respectively, for men and women, provides the respective coefficients $\{\beta_{0m}, \beta_{0f}, \beta_{km}, \beta_{kf}\}$ that are used in the OB decomposition of GWG. A novel aspect of our analysis is that the effect of the association factor (a_{ij}) can be considered in the wage equations and thus in the OB decomposition. As previously mentioned, the association factor is calculated in this study on a CT where rows (i) represent worker segments defined by the age group and gender of the employee, and columns (j) represent job segments defined by the occupation group and activity sector of the job position. Under this definition of segmentation, the estimated OB coefficient for a_{ij} lets us know who benefits most from showing a higher association (or dependence) with jobs belonging to certain occupations and activities, young or older men or women. Additionally, the formation of labour biclusters (darker areas of the heatmap) allows a segmented analysis of the wage differential. Indeed, we can analyse the GWG in those segments of the labour market where women or men (or both) tend to go (given their preferences in the labour-matching process). This segmentation analysis allows us to know both the relative situation of women in different segments of the labour market and whether their preferences in labour matching are related to going to those labour segments where the wage gap is less unfavourable for them.

3. Data description

The MCVL is a set of individual microdata extracted from the Spanish Social Security records. The Social Security information is completed with tax information from the *State Agency for Tax Administration* (*Agencia Estatal de Administración Tributaria*, AEAT) and with information from the Continuous Register provided by the *Statistics National Institute* (*Instituto Nacional de Estadística*, INE). This database offers annual information on more than a million people who appear each year in the Social Security records as recipients of income from work, subsidies or pensions. To make the sample, 4% of the

population registered in Social Security in a certain year are selected through a simple random sampling system; therefore, the MCVL is representative only of the population that is related to Social Security in the reference year – note that our analysis is cross-sectional, for the year 2021.

The MCVL is compatible with our research because it allows obtaining and estimating, cross-sectionally (for the year 2021), the wage (Mincer) equation that explains the individual's wage through a series of variables that describe attributes of the worker, the firm and the job position that he/she occupies. In the year 2021, the MCVL contains 649,893 workers, 247,822 employers (private companies or public administrations) and 1,265,406 employment episodes (labour contracts) – we only consider the employment episodes for which the information of all the variables that are going to be used in the econometric analysis is known. These episodes correspond to workers who already had a job at the beginning of the year or who found one during the year.

The wage information comes from the tax module of the MCVL, which allows us to obtain the annual income from work (whether in cash or in-kind) for each combination employee–employer (hereinafter EE) observed. Figure 3 shows the density function by gender of the annual income of the EE relationships. As can be observed, women show a higher density in intermediate wages, while men do so at higher wages. The average wage of women is €13,450 (sd = €14,214, median = €9,076), while that of men is €15,404.03 (sd = €15,551, median = €11,499). There is, therefore, an average wage gap favourable to men in the Spanish economy.

The (continuous or categorical) variables that will be used in the estimation of the wage equation and in the OB decomposition of the GWG are listed in Table A1, in the Annex. For its part, Table A2 (also located in the Annex) shows the employment distribution of those variables in Table A1 that produce the CT used to segment the job matching process; namely, the gender and age group of the worker and the occupation group and activity sector of the job position – descriptions of the rest of variables in Table A1 are available as supplementary material.

As shown in Table A2, the most frequent categories of each variable are the male gender (52.48%); the younger and intermediate age groups; the occupation groups of officers and specialists, unqualified workers and administrative assistants; and the services in the private sector, especially trade, hotels and restaurants, transport and communications and other services [2].

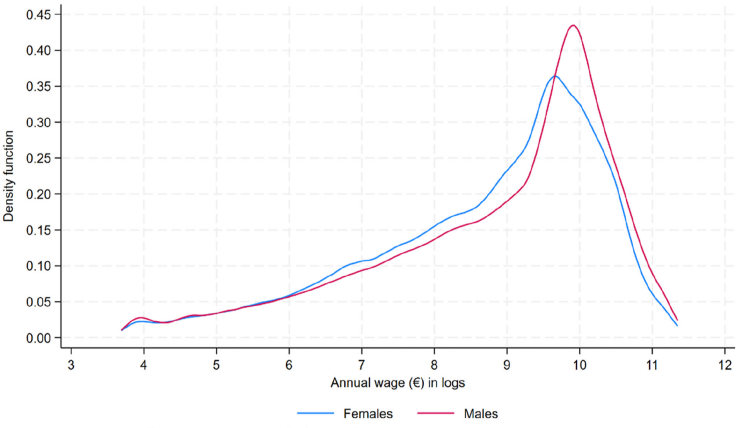


Figure 3.
Wage distribution by gender (logarithms)

Source(s): Authors' own work based on MCVL

4. Results and discussion

4.1 Results for the full sample

In this section, two econometric models are estimated, one that tries to explain the wages of workers (differentiating between men and women) and another that tries to decompose the GWG. Since the wage is not available by employment episode, but by EE combination (annual income from each payer, regardless of the number of contracts that the worker has had with that payer), the initial database of 1,265,406 employment episodes is restructured in terms of EE combinations – these combinations, 832,985 in total, constitute our sample units in the econometric models. As an EE relationship may have given rise to several employment episodes (labour contracts) in 2021, those explanatory variables directly related to the job placement are processed in the econometric models so that they are approximately representative of the annual EE relationship; these contract-dependent variables are the occupation group, the activity sector, the province of the work centre, the type of employment relationship and the type of contract. Specifically, for these five variables, we have considered in each EE relationship the category of the variable in which the worker has accumulated the longest duration during 2021 [3]. Table A3 (Annex) shows the output of the wage equation estimation (wage in logarithms). We have estimated a model for men and women considered together ("Men and Women" model in the table), a model only for men ("Men" model) and a model only for women ("Women" model). Note that, unlike other existing estimates in the literature, we have incorporated into the estimates the association factor a_{ij} {i = gender and age group; j = occupation group and sector of activity} corresponding to each EE sample unit.

Table A3 shows standard results in this area, with the best-paid workers being male, of intermediate age and longer duration in the company (in 2021), belonging to the public sector, with high educational or occupational levels and located in certain sectors of activity, such as financial and business services, supplies and extractive and manufacturing industries. Moreover, the elasticity of the annual income to the association factor of the labour segment {gender, age group; occupation, industry} is positive in the three estimates. This means that higher wages are expected for those workers who are in labour segments where workers of their age and gender group tend to be associated with the corresponding occupation group and sector of activity. Moreover, this effect is somewhat greater in the case of women (0.086 vs. 0.049), which indicates that being in a labour segment with a larger association factor generates a higher return in the women's group than in the men's group.

Table 2 shows the detailed output of the OB decomposition (Eq. (1)). The estimation shows a differential of 13 logarithmic points in favour of men – men earn, on average, 13.8% { = exp(0.13)–1} more than women – , of which 0.03 are due to the endowment differences between men and women in the different covariates considered in the model, and 0.122 points are explained by the different returns that both genders obtain from those covariates – this result is, to a certain extent, in line with the literature on Spain in this field. Note also that the effect of the interactions is −0.022 (favourable to women).

Next, we discuss the detailed results for the individual predictors. In the unexplained part of the model (coefficient component), it can be observed that men accumulate (on average) a higher duration in the same company (in the year 2021) and obtain a higher return for that duration – see Table 2. This means that given the female mean of the days worked in the company in 2021 (which enters the model in logarithmic format), the expected female wage would be 9.8% { = exp(0.09331)–1} higher if the return of that duration for women were the same as that of men (coefficient effect or unexplained part of the GWG). Similarly, the growth would be 8.8% { = exp(0.08416)–1} in the case of the variable worker's age (in logs). However, this positive effect is not observed in the association factor variable: women would earn 0.35% less if this variable were remunerated as in the case of men. Note that the value of the interaction is negligible for the duration in the company and the worker's age. In the

Table 2.
Detailed Oaxaca–Blinder decomposition

Endogenous variable: annual labour income (in logs)		Coefficient	Robust SE	z	$p > z$	[95% conf. interval]	
Full sample 832,985 EE obs.	Prediction males	8.822***	0.003	2543.2	0.000	8.815	8.829
	Prediction females	8.692***	0.003	2605.2	0.000	8.686	8.699
	Difference	0.130***	0.005	27.0	0.000	0.120	0.139
	Endowments	0.030***	0.005	6.3	0.000	0.021	0.039
	Coefficients	0.122***	0.002	73.8	0.000	0.119	0.125
	Interaction	−0.022***	0.002	−12.0	0.000	−0.026	−0.018

	Detailed decomposition	Men coefficients	Women coefficients	Mean values for men	Mean values for women	Endowments	Coefficients	Interaction
	Days working in the company (logs)	0.917***	0.897***	4.762	4.698	0.05764***	0.09331***	0.00128***
	Association factor (logs)	0.049***	0.086***	0.119	0.095	0.00206***	−0.00351***	−0.00088***
	Age (years) (logs)	0.215***	0.192***	3.654	3.657	−0.00054***	0.08416***	−0.00006**
Educational level	Basic studies	−0.054***	−0.074***	24.54%	18.87%	−0.00418***	0.00367***	0.00110***
	Higher vocational training	0.019***	−0.014***	7.07%	7.05%	0.0000	0.00233***	0.0000
	Medium vocational training	−0.005	−0.033***	6.98%	7.24%	0.00009***	0.00204***	−0.00007***
	No studies	−0.089***	−0.098***	16.62%	9.92%	−0.00659***	0.00095**	0.00064**
	School graduate	−0.067***	−0.083***	6.71%	5.45%	−0.00104***	0.00083***	0.00019***
	Other medium degrees	0.013	−0.008	0.58%	0.98%	0.0000	0.0002	−0.0001
	University diploma	0.044***	0.043***	2.45%	6.74%	−0.00185***	0.0001	0.0000
	Non-university higher studies	0.087***	0.090***	0.43%	0.53%	−0.00009***	0.0000	0.0000
	Architect or technical engineer	0.098***	0.118***	1.47%	0.62%	0.00101***	−0.00012*	−0.00017*
	High school or more	−0.008***	−0.005**	16.21%	18.57%	0.00013**	−0.0004	0.0001
	Not reported	−0.135***	−0.101***	3.37%	1.69%	−0.00170***	−0.00057***	−0.00057***
	Master or PhD	0.045***	0.090***	3.64%	6.04%	−0.00217***	−0.00274***	0.00109***
	Bachelor's degree or university degree	0.051***	0.075***	9.92%	16.30%	−0.00480***	−0.00393***	0.00154***
Occupation group	Technical engineers, graduate assistants	0.255***	0.232***	4.78%	9.55%	−0.01108***	0.00214***	−0.00107***
	Administrative officers	0.018***	0.003	8.54%	13.36%	−0.0001	0.00212***	−0.00077***
	Non-graduate assistants	0.022***	−0.024***	3.75%	3.47%	−0.00007***	0.00158***	0.00013***
	3rd officers and specialists	−0.192***	−0.198***	12.07%	9.04%	−0.00600***	0.00057*	0.00019*

(*continued*)

Table 2.

	Detailed decomposition	Men coefficients	Women coefficients	Mean values for men	Mean values for women	Endowments	Coefficients	Interaction
	Engineers, graduates and senior management	0.455***	0.451***	7.67%	8.96%	−0.00582***	0.0004	−0.0001
	Administrative assistants	−0.199***	−0.200***	7.51%	18.41%	0.02183***	0.0002	−0.0001
	Subordinates	−0.198***	−0.181***	4.32%	5.15%	0.00149***	−0.00090***	0.00014***
	Administrative and workshop managers	0.273***	0.298***	4.99%	3.81%	0.00351***	−0.00093***	−0.00029***
	1st and 2nd officers	−0.140***	−0.112***	24.64%	7.54%	−0.01916***	−0.00208***	−0.00472***
	Over 18 years unqualified or under 18 years	−0.295***	−0.269***	21.74%	20.72%	−0.00274***	−0.00535***	−0.00026***
Activity sector	Health	−0.076***	−0.145***	4.02%	15.80%	0.01704***	0.01086***	−0.00810***
	Other services	−0.004	−0.023***	23.29%	25.22%	0.00044***	0.00480***	−0.00037***
	Trade, hotels & restaurants, transport & communic	−0.033***	−0.049***	28.03%	28.30%	0.00014**	0.00453***	−0.00004**
	Education	−0.154***	−0.202***	3.51%	7.88%	0.00885***	0.00378***	−0.00210***
	Public administration	−0.074***	−0.117***	5.88%	7.98%	0.00247***	0.00351***	−0.00092***
	Extractive and manufacturing industries	0.131***	0.131***	13.45%	6.19%	0.00950***	0.0000	0.0000
	Supplies	0.166***	0.211***	1.48%	0.48%	0.00211***	−0.00022***	−0.00046***
	Construction	0.128***	0.188***	10.33%	1.61%	0.01640***	−0.00097***	−0.00524***
	Agriculture	−0.261***	−0.223***	7.82%	3.78%	−0.00903***	−0.00142***	−0.00152***
	Financial and business services	0.177***	0.230***	2.18%	2.75%	−0.00129***	−0.00144***	0.00030***
Type of contract	Permanent contract	0.188***	0.132***	51.60%	48.39%	0.00425***	0.02709***	0.00180***
	Temporary contract	0.156***	0.112***	42.49%	43.83%	−0.00150***	0.01922***	−0.00059***
	Contract for disabled workers	−0.005	0.005	0.64%	0.42%	0.0000	0.0000	0.0000
	Internships, employment promotion, and training	−0.144***	−0.128***	0.65%	0.70%	0.00007***	−0.0001	0.0000
	Not reported	−0.093**	−0.071	0.03%	0.02%	0.0000	0.0000	0.0000
	Public sector workers	0.391***	0.3***	5.26%	11.17%	−0.0177***	0.0102***	−0.0054***
	Constant	4.026***	4.199***	1.000	1.000		−0.17299***	

Note(s): *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$. Standard errors adjusted for 640,214 clusters of individuals (workers). Other control variables: Worker attributes (Nationality, Social benefits received in 2021, Income received in 2021 from professional activities, Number of labour contracts with the company during 2021, Duration since the first contract with the company (years), Collective agreement); Job placement attributes (Province, Percentage of the income from work that is in kind); Firm attributes (Legal person vs. natural person, Number of workers (logs)). The complete table is offered as supplementary material (available online)

Source(s): Authors' own work based on MCVL

coefficients component, we also find dummy categories for which the average wage of women would increase if the men's coefficients [4] were applied to the women's characteristics, for instance: the health sector (1.1%), the permanent (2.75%) or temporary (1.94%) contract and the public sector category (1.02%). Finally, for the model intercept (−15.9% = exp(0.17299)–1), the average wage of women would decrease if the men's constant were applied to women – in other words, when it comes to the constant of the model, being a woman contributes to reducing the wage gap. The case of the model constant is interesting since it includes the effects of unobservable variables not taken into account (i.e. not included in the model).

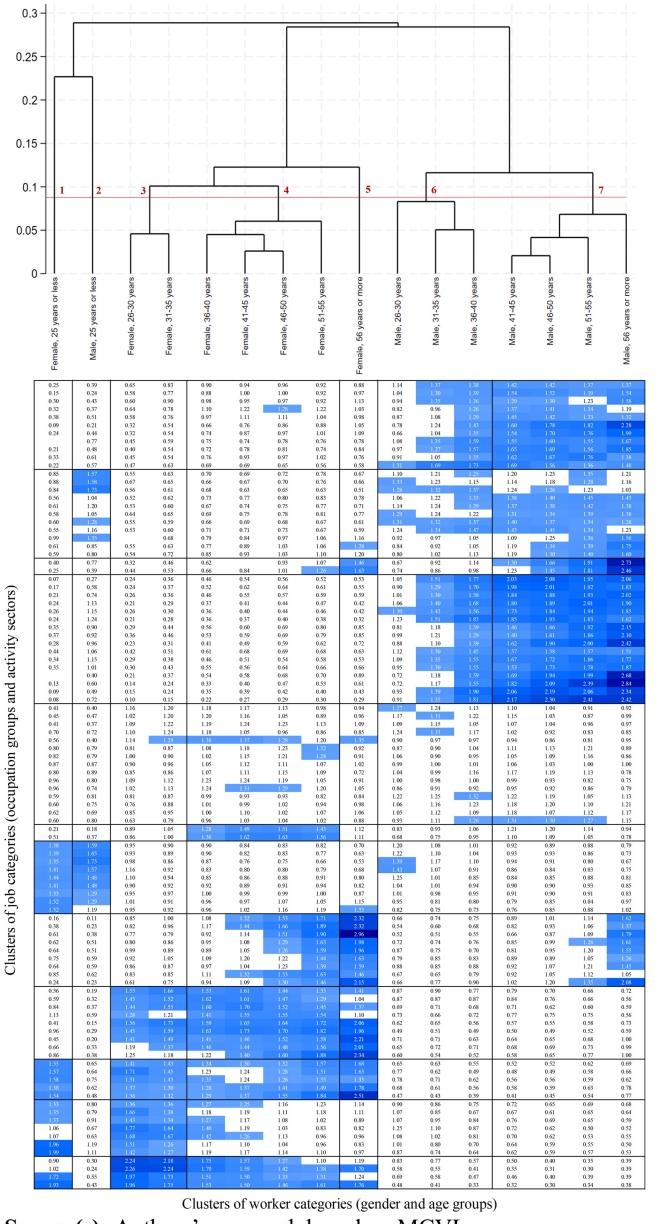
As for the explained part of the model (endowments component), we observe that, given the female return of the days worked in the company (in 2021), the expected female wage would be 5.9% { = exp(0.05764)–1} higher if the average level of that duration for women were the same as that of men. Likewise, a small but significant positive contribution to the wage gap is observed in the association factor variable (0.21%). Moreover, for some dummy categories, the average wage of women would increase if they had the same average characteristics as men; some of these categories are the occupation group of administrative assistants (salary 2.2% higher) and the activity sectors of Health (1.7%) and Construction (1.65%). For example, in the health sector, women have a greater representation than men (men 4.02% vs women 15.8%) but belonging to this sector penalises them compared to other sectors; therefore, the component $\widehat{\beta}_{\text{health,f}} \cdot (\overline{X}_{\text{health,m}} - \overline{X}_{\text{health,f}})$ turns out to be positive –note that we are simulating that women lose weight $(\overline{X}_{\text{health,m}} - \overline{X}_{\text{health,f}} < 0)$ in a sector that penalises them in wages $(\widehat{\beta}_{\text{health,f}} < 0)$. On the contrary, some dummy categories for which the average wage of women would decrease if they had the same average characteristics as men are the public sector category (wage 1.76% lower) and the occupation groups technical engineers and graduate assistants (wage 1.1% lower) and 1st and 2nd officers (wage 1.9% lower). In the case of the public sector category and the first occupation group, because women are more represented than men and the return of the corresponding category is positive for them (women). In the second occupation group (1st and 2nd officers), because men are more represented than women, but women obtain a negative return for belonging to this group. Consequently, in these three categories (public sector and the two occupation groups pointed out), being a woman contributes to reducing the GWG.

4.2 Results by biclusters

(1) *Job placement database*

In this section, the Spanish labour market is analysed considering that it can be endogenously divided into labour segments (biclusters) that are based on how workers, classified according to their gender and age group, match up with jobs, classified according to their occupation group and activity sector. Using the methodology described in Section 2 on the database of 1,265,406 job placements, we obtain the association factors table (or heatmap) of Figure 4.

The columns of the figure represent 16 crossed categories of workers (worker segments) that come from combining 2 genders and 8 age groups {25 years or less, 26–30 years, 31–35, 36–40, 41–45, 46–50, 51–55 and 56 years or more}, while the rows represent 100 crossed categories of jobs (job segments) that come from combining 10 occupation groups and 10 activity sectors – the categories of occupations and activities are shown in Table 3. Both rows and columns have been, respectively, clustered to have an orderly view of the labour market. In Figure 4, we only show the column dendrogram, since the row dendrogram is too large (100 job segments). In addition, to better interpret the heatmap, (1) we have coloured the cells blue, with an association factor greater than one and (2) the higher the association factor of a cell with $a_{ij} > 1$, a darker shade of blue.



Source(s): Authors' own work based on MCVL

Figure 4.
Heatmap of the Spanish labour market

Figure 4 allows us to reach conclusions of interest. The labour market to which women tend to go is visibly different from that of men. The column dendrogram (by gender and age group) shows that men and women are differently matched to different occupation and activity groups – note that this dendrogram groups first by gender and then, within each gender, by age group. Moreover, the groups of men and women of 25 years or younger have little

		Females ($a_{ij} \geq 1.25$)	Males ($a_{ij} \geq 1.25$)	Females (full sample)	Males (full sample)
Activity sector	Agriculture	0.01%	4.4%	1.6%	3.7%
	Extractive and manufacturing industries	0.01%	9.8%	2.3%	5.5%
	Utilities (water, electricity, gas)	0.0%	1.4%	0.2%	0.7%
	Construction	0.0%	9.7%	0.6%	4.5%
	Financial and business services	1.3%	0.5%	1.0%	1.0%
	Trade, hotels and restaurants, transport and communications	8.1%	10.2%	12.1%	13.1%
	Health	20.7%	0.1%	8.9%	2.7%
	Education	7.3%	0.0%	3.6%	1.8%
	Public administration	5.8%	1.7%	3.2%	2.5%
	Other services	6.9%	12.3%	14.0%	17.0%
Occupation group	Engineers, graduates and senior management	6.0%	1.6%	3.8%	3.7%
	Technical engineers, graduate assistants	9.3%	0.4%	4.7%	2.4%
	Administrative and workshop managers	0.7%	1.9%	1.9%	3.0%
	Administrative officers	4.6%	1.0%	5.0%	3.8%
	Non-graduate assistants	1.4%	1.4%	1.5%	1.7%
	1st and 2nd officers	0.3%	23.7%	3.8%	11.5%
	Subordinates	5.3%	0.6%	2.9%	2.2%
	Administrative assistants	13.8%	0.2%	8.2%	3.8%
	3rd officers and specialists	3.3%	7.5%	4.8%	6.9%
	Over 18 years unqualified or under 18 years	5.5%	11.8%	11.1%	13.5%

Source(s): Authors' own work based on MCVL

Table 3. Employment distribution of the cells with the highest association factor

similarity with the rest of the groups and show a relatively high dissimilarity between them. Therefore, we can then deduce that there are significant gender differences in how young workers approach the labour market. Another difference between men and women is observed in the age group between 36 and 40 years. This group is initially arranged with the group of 31–35 years in the case of men and with the cluster of 41–50 in the case of women, giving the impression that age stigmatises women before men.

Table 3 contains information about the rows (occupations and industries) of the heatmap from a gender perspective. In the table, we show the distribution of job placements (by gender and occupation group and gender and activity sector) of those placements that belong to the cells of the heatmap corresponding to the highest a_{ij} quartile (Q1) for both men and women – i.e. the cells in Figure 4 with a more intense blue colour. In both cases (men and women), the top 75% of a_{ij} distribution is reached approximately when this factor exceeds the value of 1.25. For comparative purposes, we also show in the table the job placement distributions for the entire sample.

Regarding the sectors of activity, the Q1-zone of the heatmap is very different for men and women. While women tend to be associated with service activities (especially health), men are associated mainly with extractive, manufacturing and construction industries and with services more typical of the private sector of the economy such as trade, hospitality, transport, communications and others. The women's job placements in health, education and public administration account for 33.7% of the job placements in the Q1-zone of the heatmap, this percentage being 15.7% if we analyse the entire heatmap. Interestingly, the finance and business service sectors are not a major focus of job attraction for men and women; in the matching game, they better prefer other sectors of activity. As for the

occupation groups, the differences by gender in the Q1-zone of the heatmap are also notable. The women's job placements in the first four occupational groups (those with the highest qualification) exceed 20% in the Q1-zone (this percentage is 15.4% in the full heatmap), while the men's job placements do not reach 5% in these four groups (this percentage is approximately 13% for men in the full heatmap). Also noteworthy is the high relative weight of 1st, 2nd and 3rd officers in the case of men; in the darkest areas of the heatmap, male officers represent more than 30% of the job placements (18.4% if we look at the entire heatmap).

(2) *Employee–Employer database*

The previous heatmap (Figure 4) allows us to observe the existence of labour market segments (biclusters) that can be analysed from the perspective of the GWG. To do this, we need to use the EE database, because it is the one that allows us to use the annual earnings from each EE relationship. As can be seen, the heatmap in Figure 4 has been divided into 7-column clusters and 12-row clusters. This makes 84 cluster intersections of which we have selected the 21 that show a higher degree of association between their respective rows and columns. We refer to these 21 cluster crossings as biclusters or labour market segments, which can be analysed using the OB decomposition.

Figure 5 relates, for each bicluster, the GWG and the explained and unexplained components (in percentage) of the OB estimate. Figure 5(a) shows that a larger wage gap is observed in those biclusters with a greater relative weight of the explained part of the gap; the opposite happens with the unexplained part of the model – Figure 5(b). These results would indicate that in those biclusters where wage differences are more important, these differences are mainly due to the characteristics of men and women and not so much to the different return of these characteristics.

Our microdata allows an in-depth analysis of any bicluster of interest. As an example, we describe in Table 4 the three biclusters that appear labelled in Figures 5 and 6. These are biclusters with a significant volume of job placements, one where women show a relatively high a_{ij} factor, another where this happens to men and a third one where this happens to both genders.

The bicluster where women show a relatively high association factor (average women $a_{ij} = 1.48$, average men $a_{ij} = 0.97$) has 20,658 women's job placements and 14,009 men's job placements – this bicluster is labelled as "Women show higher a_{ij}" in Figures 5 and 6. We are talking about workers between 26 and 35 years old, administrative assistants, technical engineers or graduate assistants in the private service sector (commerce, hospitality, transport, communications and other services). In this bicluster, the wage gap is 0.18 logarithmic points (in favour of men) and is explained in almost equal parts by endowments and coefficients. For its part, the bicluster where men show a relatively high association factor (labelled as "Men show higher a_{ij}" in Figures 5 and 6; average women $a_{ij} = 0.8$, average men $a_{ij} = 1.33$) is made up of workers aged 41 or older and has 32,110 job placements for women and 71,070 for men. This bicluster covers private service activities (like the previous ones), agriculture and extractive and manufacturing industries (it is quite transversal) and corresponds to officers or low-skilled workers. In this bicluster, the wage gap is 0.43 logarithmic points (in favour of men) and is mainly due to the characteristics of each one, although the unexplained part of the gap is also important. Finally, the bicluster where men and women show high association factors (label "Both show high a_{ij}" in Figures 5 and 6; average women $a_{ij} = 1.42$, average men $a_{ij} = 1.48$) is made up of workers aged 25 or less (31,682 job placements for women and 44,577 for men) with low qualifications (over 18 years unqualified or under 18 years, or 3rd officers and specialists) in the private service sector. In this bicluster, the wage gap is 0.26 logarithmic points (in favour of men) and is mainly due to the characteristics/endowments of each one.

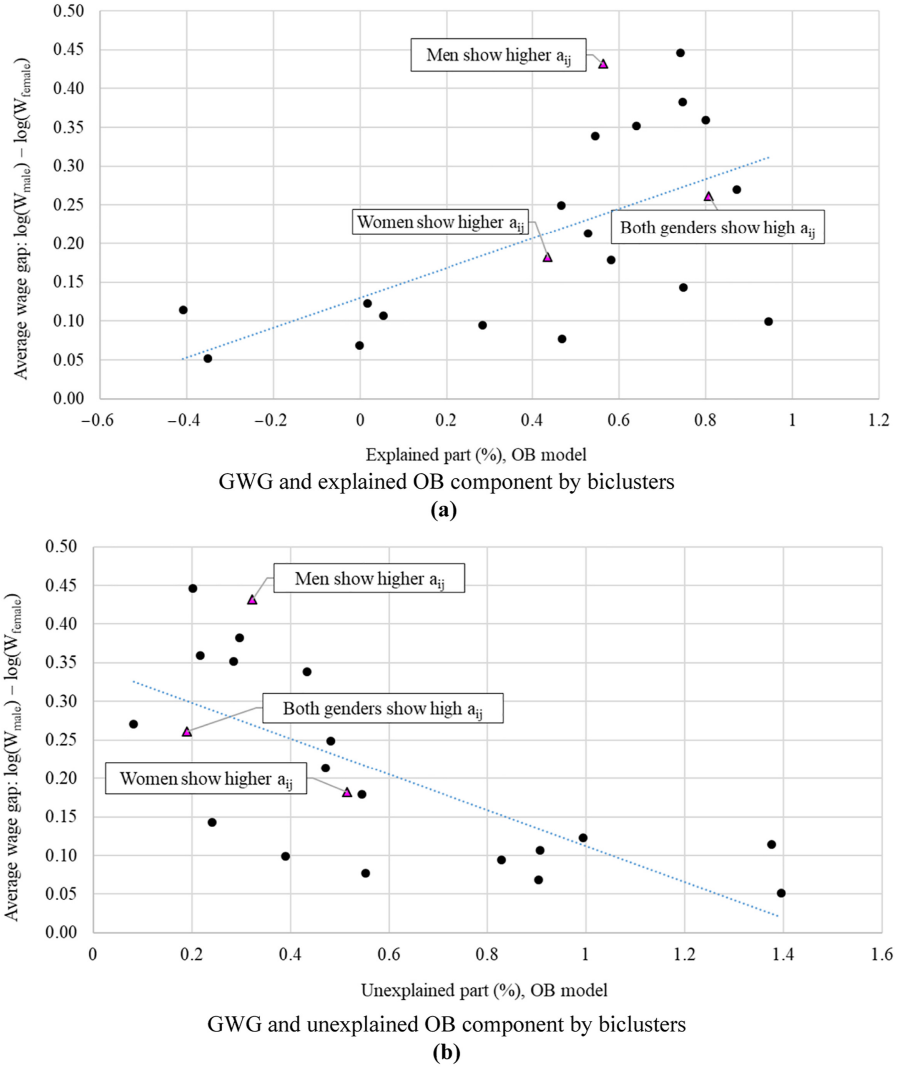


Figure 5.
GWG gap and OB components by biclusters

Source(s): Authors' own work based on MCVL

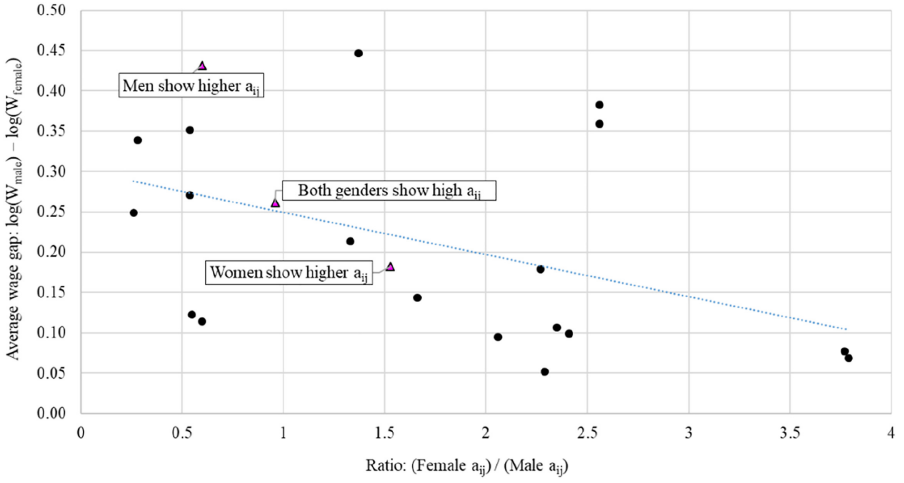
Obviously, we cannot think of applying common labour policies to labour market segments with such different characteristics, wage gaps and gap decompositions. For example, while in the bicluster where men and women both show a high association factor (bicluster of young workers), it would be necessary to investigate why the endowments are so favourable to men in terms of remuneration; in the other two biclusters, it would also be necessary to investigate what factors explain that women obtain lower returns than men for their contribution to the productive activity.

Our segmentation analysis ends by relating the average GWG of each bicluster (wage difference in logs) to the ratio of the respective average association factors of women and men in

Bicluster	Results	Coefficient	Robust SE	z	P > z	[95% conf. interval]	
Ratio female a_{ij}/male $a_{ij} = 1.53$	Prediction males	8.485***	0.012	827.2	0.000	9.855	9.902
	Prediction females	8.303***	0.008	1,152.6	0.000	9.755	9.788
	Difference	0.182***	0.015	7.3	0.000	0.078	0.136
	Endowments	0.080***	0.014	0.4	0.682	−0.022	0.033
	Coefficients	0.094***	0.005	17.9	0.000	0.086	0.108
	Interaction	0.009**	0.003	1.4	0.166	−0.002	0.010
Ratio female a_{ij}/male $a_{ij} = 0.6$	Prediction males	8.747***	0.009	969.0	0.000	8.729	8.765
	Prediction females	8.316***	0.012	696.8	0.000	8.292	8.339
	Difference	0.431***	0.015	28.8	0.000	0.402	0.461
	Endowments	0.243***	0.014	17.3	0.000	0.216	0.271
	Coefficients	0.139***	0.005	26.1	0.000	0.129	0.150
	Interaction	0.049***	0.004	13.1	0.000	0.042	0.056
Ratio female a_{ij}/male $a_{ij} = 0.96$	Prediction males	7.321***	0.011	649.7	0.000	7.299	7.343
	Prediction females	7.060***	0.012	573.2	0.000	7.036	7.085
	Difference	0.261***	0.017	15.6	0.000	0.228	0.294
	Endowments	0.211***	0.016	13.3	0.000	0.180	0.242
	Coefficients	0.049***	0.005	9.8	0.000	0.040	0.060
	Interaction	0.001	0.003	0.2	0.865	−0.005	0.006

Source(s): Authors' own work based on MCVL

Table 4.
OB estimation for the largest biclusters of each type



Source(s): Authors' own work based on MCVL

Figure 6.
GWG and relative association factors by biclusters

the corresponding bicluster [5] (Figure 6). As can be seen in the figure, as the relative association factor of women increases, the wage differential narrows – the high dispersion of the relationship is reasonable if we consider that the idiosyncratic labour markets in the figure can be very different. This negative trend means that women tend to be placed in those labour biclusters where the wage differentials with men are smaller. For them, not only is the wage level important but also the situation of wage inequality with respect to the male workers.

5. Conclusions

In this study, we have tried to show that the process of labour matching and the possible existence of gender wage discrimination in the Spanish labour market are phenomena that cannot be studied by treating the whole population/sample (employment episodes) homogeneously. On the contrary, it is necessary to segment the employment to find idiosyncratic labour markets that can be analysed from a gender perspective. Using matched EE data for the year 2021 in Spain (we use the MCVL provided by the Spanish Social Security), we take four variables that are key to segmenting the labour matching process: the gender and age group of the worker and the occupation group and activity sector of the job placement. By applying CT and hierarchical clustering techniques, we create a heatmap (clustered association table) of employment episodes in the year 2021 that allows us to identify employment biclusters (idiosyncratic labour markets) where workers of one gender (men or women) show a higher degree of association than workers of the other gender with certain sectors of activity and occupation groups and vice versa; the analysis is further refined by discriminating workers by different age groups.

Our study provides an in-depth analysis of workers' remuneration in the Spanish economy. We estimate a wage equation for the full sample and for men and women separately and perform the OB decomposition to try to explain the existing wage differential in favour of men. The OB estimation shows a wage differential of 13 logarithmic points in favour of men, most of it due to the different returns that both genders obtain from their respective matching-related attributes – this result is in line with the literature on Spain in this field. Additionally, we analyse the GWG in different idiosyncratic markets which are defined by different clusters of occupation groups and activity sectors where certain clusters of men and/or women (of different age groups) tend to seek employment and get a job. These crossings of clusters with high internal association (denoted as biclusters) are extracted from the mentioned employment heatmap. The application of the OB model to these idiosyncratic markets based on "who matches with whom" constitutes a novel aspect within the literature on gender wage discrimination.

Our gender labour segmentation analysis shows that the labour market to which women attend is visibly different from that of men. For instance, while women tend to be associated with service activities (especially, public services, such as health or education), men are mainly associated with extractive, manufacturing and construction industries and with private services. Furthermore, it is observed that women are associated more intensely than men with the highest occupation groups (those with the highest qualification). This segmented scenario implies that the phenomenon of gender wage discrimination cannot be analysed with a global and homogeneous vision of the labour market and addressed with general policies. Effectively, the different idiosyncratic markets extracted from our employment heatmap show different wage gaps, as well as different weights of the observed (endowments) and unobserved (coefficients) heterogeneity. The overall analysis of these labour market segments produces two interesting conclusions. On the one hand, women tend to be placed in those labour biclusters where the wage differentials with men are smaller. For women, not only is the wage level important, but also the situation of wage inequality with respect to male workers. On the other hand, in those biclusters where wage differences are more important, these differences are mainly due to the characteristics of men and women, and not so much to the different return of these characteristics, although this last component is not negligible. In fact, when we use data from the entire labour market, which would imply using information from the entire heatmap (and not just from the idiosyncratic biclusters), the unexplained component of the OB decomposition explains most of the wage gap.

Our combination of methodologies (clustered contingency tables and wage gap decomposition) is versatile and flexible (it can be applied to other economies or worker

groups: migrants/natives, public/private employees, university graduates/other educational levels, etc.) and provides a better understanding of the underlying segmentation in the labour matching process and the effect of this segmentation in the gender wage issue. A more comprehensive knowledge of the underlying structure of the labour market helps in the efficient design of labour policies from a gender perspective; for instance, it would be necessary to act in those idiosyncratic labour markets where men continue to earn more than women fundamentally for reasons that are not justified by their respective characteristics. Our methodology allows us to identify these markets. The quantile analysis of the wage differential by labour segments, the consideration of the regional dimension in the heatmap and the application of our methodological tools to specific groups in the labour market are other possible lines of extension of our research.

Notes

1. The concept of association factor was introduced by Good (1956).
2. This last sector includes activities such as professional, scientific and technical activities, administrative and support service activities, recreational, cultural and sports activities, real estate activities, computer activities, associative activities and activities of households as employers or producers (self-consumption).
3. We have also estimated the models considering the two longest duration categories of each job-dependent variable, but we have discarded these estimates because the improvement in model fit is small and more degrees of freedom are lost in the estimation.
4. Observe in Table 2 that the OB model transforms the coefficients of the dummy variables so that they reflect deviations from the "grand mean" (in other words, the modified coefficients will sum up to zero over all categories) rather than deviations from the reference category. This deviation contrast transformation allows the model output to be invariant to the choice of the (omitted) base category. On this transformation, see for example Yun (2005).
5. To obtain the average association factor for each bicluster, we have created an auxiliary 7×12 CT where each cell represents the job placements for the crossing of the corresponding row cluster and column cluster. Our methodology of associations can be applied to this more compact CT to obtain the association factor between each cluster crossing.

References

Agresti, A. (2013), *Categorical Data Analysis*, 3rd ed., Wiley, New York.

Álvarez de Toledo, P., Núñez, F. and Usabiaga, C. (2018), "Matching and clustering in square contingency tables. Who matches with whom in the Spanish labour market", *Computational Statistics and Data Analysis*, Vol. 127, pp. 135-159, doi: 10.1016/j.csda.2018.05.012.

Álvarez de Toledo, P., Núñez, F. and Usabiaga, C. (2020), "Matching in segmented labor markets: an analytical proposal based on high-dimensional contingency tables", *Economic Modelling*, Vol. 93, pp. 175-186, doi: 10.1016/j.econmod.2020.07.019.

Anghel, B., Lacuesta, A. and Regil, A. (2020), "Transferability of workers' skills in sectors potentially affected by covid-19", Analytical Articles, *Economic Bulletin*, Banco de España, Article 2/2020, pp. 1-13.

Antón, J.I. and Muñoz de Bustillo, R. (2015), "Public-private sector wage differentials in Spain: an updated picture in the midst of the Great Recession", *Investigación Económica*, Vol. 74 No. 292, pp. 115-157, doi: 10.1016/j.inveco.2015.08.005.

Arrazola, M., De Hevia, J., Perrote, I. and Sánchez, R. (2022), "Brecha de género en la inserción laboral de los graduados españoles", *Papeles de Trabajo del Instituto de Estudios Fiscales, Serie Economía*, Vol. 10, pp. 1-69.

Autor, D.H., Katz, L.F., Kearney, M.S. (2006), "Rising wage inequality: the role of composition and prices", NBER Working Papers, No. 11628.

Autor, D.H., Katz, L.F. and Kearney, M.S. (2008), "Trends in U.S. wage inequality: revising the revisionists", *The Review of Economics and Statistics*, Vol. 2 No. 90, pp. 300-323, doi: 10.1162/rest.90.2.300.

Barth, E. and Dale-Olsen, H. (2009), "Monopsonistic discrimination and the gender-wage gap", *Labour Economics*, Vol. 16 No. 5, pp. 589-597, doi: 10.1016/j.labeco.2009.02.004.

Becker, G. (1971), *The Economics of Discrimination*, 2nd ed., University of Chicago Press, Chicago (Il.).

Bisello, M., Maccarrone, V. and Fernández-Macías, E. (2022), "Occupational mobility, employment transitions and job quality in Europe: the impact of the Great Recession", *Economic and Industrial Democracy*, Vol. 43 No. 2, pp. 585-611, doi: 10.1177/0143831x20931936.

Blinder, A. (1973), "Wage discrimination: reduced form and structural estimates", *Journal of Human Resources*, Vol. 8 No. 4, pp. 436-455, doi: 10.2307/144855.

Cahuc, P., Carcillo, S. and Zylberberg, A. (2014), *Labor Economics*, 2nd ed., MIT press, Boston (Mass.).

Caparrós-Ruiz, A. (2016), "Wage growth and occupational mobility in Spain: movers vs stayers", *International Journal of Social Economics*, Vol. 43 No. 12, pp. 1481-1506, doi: 10.1108/ijse-03-2015-0071.

Couceiro de León, A. and Dolado, J.J. (2023), "Revisiting the public-private wage gap in Spain: new evidence and interpretation", *SERIEs*, Vol. 14 Nos 3-4, pp. 353-377, doi: 10.1007/s13209-023-00277-z.

Dueñas, D. and Moreno, A. (2018), "Descomposición del GAP salarial por género en España, Francia y Alemania", *Investigación y género. Reflexiones desde la investigación para avanzar en igualdad, VII Congreso Universitario Internacional Investigación y Género*, SIEMUS, pp. 147-168, (2018).

EU-SILC (2016), *Survey of Income and Living Conditions (SILC)*, Statistical Release, Central Statistics Office (CSO), Dublin.

Fernández-Cerezo, A. and Montero, J.M. (2021), *A Sectoral Analysis of the Future Challenges Facing the Spanish Economy*, Documentos Ocasionales, Banco de España, Madrid, No. 2133.

Firpo, S., Fortin, N. and Lemieux, T. (2018), "Decomposing wage distributions using recentered influence function regressions", *Econometrics*, Vol. 6 No. 2, pp. 1-40, doi: 10.3390/econometrics6020028.

Fortin, N., Lemieux, T. and Firpo, S. (2011), "Decomposition methods in economics", in Orley, A. and David, C. (Eds), *Handbook of Labor Economics*, North-Holland, Amsterdam, Vol. IV.A, pp. 1-102.

Gale, D. and Shapley, L.S. (1962), "College admissions and the stability of marriage", *The American Mathematical Monthly*, Vol. 69 No. 1, pp. 9-15, doi: 10.1080/00029890.1962.11989827.

García, I. and Molina, J.A. (2002), "Inter-regional wage differentials in Spain", *Applied Economics Letters*, Vol. 9 No. 4, pp. 209-215, doi: 10.1080/13504850110065849.

Good, I.J. (1956), "On the estimation of small frequencies in contingency tables", *Journal of the Royal Statistical Society: Series B*, Vol. 18 No. 1, pp. 113-124, doi: 10.1111/j.2517-6161.1956.tb00216.x.

Gordon, N. and Morton, T. (1974), "A low mobility model of wage discrimination with special reference to sex differential", *Journal of Economic Theory*, Vol. 7 No. 3, pp. 241-253, doi: 10.1016/0022-0531(74)90095-7.

Guner, N., Kaya, E. and Sánchez-Marcos, V. (2014), "Gender gaps in Spain: policies and outcomes over the last three decades", *SERIEs*, Vol. 5 No. 1, pp. 61-103, doi: 10.1007/s13209-014-0104-z.

Heckman, J. (1979), "Sample selection bias as a specification error", *Econometrica*, Vol. 47 No. 1, pp. 153-161, doi: 10.2307/1912352.

Hidalgo, M. (2010), "Wage inequality in Spain, 1980-2000: the case of male head-of-household", *Estadística Española*, Vol. 52 No. 174, pp. 333-366.

Hospido, L. and Moral-Benito, E. (2016), "The public sector wage premium in Spain: evidence from longitudinal administrative data", *Labour Economics*, Vol. 42, pp. 101-122, doi: 10.1016/j.labeco.2016.08.001.

Jann, B. (2008), "The Oaxaca-Blinder decomposition for linear regression models", *The Stata Journal*, Vol. 8 No. 4, pp. 435-479.

Juhn, C., Murphy, K. and Pierce, B. (1991), "Accounting for the slowdown in black-white wage convergence", in Kosters, M. (Ed.), *Workers and Their Wages*, AEI Press, Washington, DC, pp. 107-143.

Juhn, C., Murphy, K. and Pierce, B. (1993), "Wage inequality and the rise in returns to skill", *Journal of Political Economy*, Vol. 101 No. 3, pp. 410-442, doi: 10.1086/261881.

Kim, C. (2010), "Decomposing the change in the wage gap between white and black men over time, 1980-2005: an extension of the Blinder-Oaxaca decomposition method", *Sociological Methods and Research*, Vol. 38 No. 4, pp. 619-651, doi: 10.1177/0049124110366235.

Kröger, H. and Hartmann, J. (2021), "Extending the Kitagawa-Oaxaca-Blinder decomposition approach to panel data", *The Stata Journal*, Vol. 21 No. 2, pp. 360-410, doi: 10.1177/1536867x211025800.

Machado, J. and Mata, J. (2005), "Counterfactual decomposition of changes in wage distributions using quantile regression", *Journal of Applied Econometrics*, Vol. 20 No. 4, pp. 445-465, doi: 10.1002/jae.788.

Manning, A. and Petrongolo, B. (2017), "How local are labor markets? Evidence from a spatial job search model", *American Economic Review*, Vol. 107 No. 10, pp. 2877-2907, doi: 10.1257/aer.20131026.

Melly, B. (2005), "Decomposition of differences in distribution using quantile regression", *Labour Economics*, Vol. 12 No. 4, pp. 577-590, doi: 10.1016/j.labeco.2005.05.006.

Mosteller, F. (1968), "Association and estimation in contingency tables", *Journal of the American Statistical Association*, Vol. 63 No. 321, pp. 1-28, doi: 10.2307/2283825.

Murillo-Huertas, I.P., Ramos, R. and Simón, H. (2017), "Regional differences in the gender wage gap in Spain", *Social Indicators Research*, Vol. 134 No. 3, pp. 981-1008, doi: 10.1007/s11205-016-1461-8.

Murillo-Huertas, I.P., Ramos, R. and Simón, H. (2020), "Revisiting interregional wage differentials: new evidence from Spain with matched employer-employee data", *Journal of Regional Science*, Vol. 60 No. 2, pp. 296-347, doi: 10.1111/jors.12459.

Ñopo, H. (2008), "Matching as a tool to decompose wage gaps", *Review of Economics and Statistics*, Vol. 90 No. 2, pp. 290-299, doi: 10.1162/rest.90.2.290.

Oaxaca, R. (1973), "Male-female wage differentials in urban labor markets", *International Economic Review*, Vol. 14 No. 3, pp. 693-709, doi: 10.2307/2525981.

Roth, A.E. and Sotomayor, M. (1992), "Two-sided matching", in Aumann, R. and Hart, S. (Eds), *Handbook of Game Theory with Economic Applications*, North-Holland, Amsterdam, pp. 485-541, 1, Ch.16.

Smith, J.P. and Welch, F.R. (1989), "Black economic progress after Myrdal", *Journal of Economic Literature*, Vol. 27 No. 2, pp. 519-564.

Yun, M.S. (2005), "A simple solution to the identification problem in detailed wage decompositions", *Economic Inquiry*, Vol. 43 No. 4, pp. 766-772, doi: 10.1093/ei/cbi053.

Corresponding author
Carlos Usabiaga can be contacted at: cusaiba@upo.es

Supplementary material
The supplementary material for this article can be found online.

Annex

	Explanatory variables	Comments or clarifications
Worker attributes	Gender	0: Female; 1: Male
	Age (in years)	Age of the worker in mid-2021 (obtained from the date of birth)
	Nationality	Each nationality is assigned a three-position numerical code
	Educational level	Coding: CNED-2014 (National Education Classification, source: INE)
	Social benefits received in 2021	Mainly unemployment benefits from the Social Security
	Income received in 2021 from professional activities	Mainly income from professional and farming activities
	Cumulative duration in the company in 2021 (days)	Accumulating the duration of all contracts with the company in 2021 (weighting each contract duration by its part-time coefficient)
	Number of labour contracts with the company during 2021	The annual employee–employer relationship can give rise to several contracts
	Duration since the first contract with the company (years)	This data can be obtained because the entire working life of the worker is observed
	Collective agreement	Identifies the scope of the collective agreement in force for employees (if any)
Job placement attributes	Occupation group	It is the contribution group to SS assigned to the worker. Coding: CNO-2011 (National Classification of Occupations, 2011; source: INE)
	Activity sector	Coding: CNAE-2009 (National Classification of Economic Activities, 2009; source: INE)
	Province of the work centre	52 NUTS-3 regions
	Type of Employment	Identifies certain groups of affiliates with contribution peculiarities
	Relationship of the Contribution Account	
	Type of contract	Identifies the type of contract between the employee and the employer
	Percentage of the income from work that is in kind	This percentage can be zero
Employer attributes	Company size	Number of workers
	Legal form of the company	Form or legal personality of the employer according to the classification established by the State Agency for Tax Administration

Table A1. Explanatory variables of the wage econometric models

Source(s): Authors' own work based on MCVL

		Frequency	Percentage
Gender	Female	601,315	47.52
	Male	664,091	52.48
Age	25 years or less	189,236	14.95
group	26-30 years	165,500	13.08
	31-35 years	153,794	12.15
	36-40 years	155,614	12.3
	41-45 years	174,320	13.78
	46-50 years	158,508	12.53
	51-55 years	126,879	10.03
	56 years o more	141,555	11.19
Occupation group	Engineers, graduates and senior management	94,879	7.5
	Technical engineers, graduate assistants	88,928	7.03
	Administrative and workshop managers	61,923	4.89
	Non-graduate assistants	40,699	3.22
	Administrative officers	111,347	8.8
	Subordinates	64,206	5.07
	Administrative assistants	152,200	12.03
	1st and 2nd officers	193,091	15.26
	3rd officers and specialists	147,371	11.65
	Over 18 years unqualified or under 18 years	310,762	24.56
Activity sector	Agriculture	67,273	5.32
	Extractive and manufacturing industries	97,932	7.74
	Supplies	11,364	0.9
	Construction	64,104	5.07
	Trade, hotels & restaurants, transport & communic.	319,245	25.23
	Financial and business services	24,688	1.95
	Public administration	72,667	5.74
	Education	68,722	5.43
	Health	146,715	11.59
	Other services	392,500	31.02
	Extraterritorial organisations	196	0.02

Source(s): Authors' own work based on MCVL

Table A2. Employment distribution by characteristics of workers and jobs

Table A3.
Wage equation estimation (wage in logarithms)

Endogenous variable: Annual labour income (in logs)		Men & Women 832,985 EE obs. R^2 adj. 91.1% SE adjusted for 640,214 clusters of workers Robust			Men 445,902 EE obs. R^2 adj. 91.2% SE adjusted for 337,629 clusters of workers Robust			Women 387,083 EE obs. R^2 adj. 91% SE adjusted for 302,585 clusters of workers Robust		
		Coefficient	SE	t	Coefficient	SE	t	Coefficient	SE	t
	Gender	0.116***	0.001	83.3						
	Days working in the company (logs)	0.907***	0.001	1,081.3	0.917***	0.001	819.1	0.897***	0.001	707.6
	Age (years) (logs)	0.201***	0.002	82.2	0.215***	0.003	61.8	0.192***	0.003	55.9
	Association factor (logs)	0.047***	0.001	31.5	0.049***	0.003	14.6	0.086***	0.003	25.0
Educational level (reference: Master or PhD)	Architect or technical engineer	0.042***	0.006	7.6	0.053***	0.007	7.6	0.027**	0.011	2.6
	Non-university higher studies	0.017	0.012	1.4	0.042**	0.018	2.4	−0.0003	0.016	−0.02
	Bachelor's degree or university degree	−0.006**	0.003	−2.0	0.006	0.005	1.2	−0.015***	0.004	−3.7
	University diploma	−0.035***	0.004	−8.1	−0.001	0.008	−0.1	−0.047***	0.005	−9.5
	Higher vocational training	−0.066***	0.004	−17.6	−0.026***	0.006	−4.6	−0.104***	0.005	−21.1
	Other medium degrees	−0.072***	0.008	−9.4	−0.032**	0.012	−2.6	−0.098***	0.010	−10.1
	High school or more	−0.078***	0.003	−22.6	−0.053***	0.005	−9.9	−0.096***	0.004	−21.3
	Medium vocational training	−0.089***	0.004	−23.5	−0.050***	0.006	−8.7	−0.123***	0.005	−24.4
	Basic studies	−0.135***	0.004	−38.3	−0.099***	0.005	−18.5	−0.164***	0.005	−35.3
	School graduate	−0.146***	0.004	−35.1	−0.112***	0.006	−18.6	−0.173***	0.006	−29.6
	No studies	−0.168***	0.004	−44.5	−0.134***	0.006	−23.8	−0.189***	0.005	−36.6
	Not reported	−0.205***	0.007	−30.3	−0.180***	0.009	−20.1	−0.191***	0.011	−18.1
Occupation group (reference: Engineers, graduates and senior management)	Administrative and workshop managers	−0.170***	0.004	−39.7	−0.182***	0.006	−31.4	−0.153***	0.006	−23.6
	Technical engineers, graduate assistants	−0.205***	0.003	−73.8	−0.200***	0.004	−44.9	−0.218***	0.004	−60.1
	Administrative officers	−0.447***	0.003	−143.6	−0.437***	0.005	−90.1	−0.448***	0.004	−111.1
	Non-graduate assistants	−0.452***	0.004	−112.2	−0.433***	0.006	−77.5	−0.474***	0.006	−80.8
	1st and 2nd officers	−0.582***	0.003	−183.3	−0.595***	0.005	−130.2	−0.563***	0.005	−107.2

(*continued*)

Table A3.

Endogenous variable: Annual labour income (in logs)		Men & Women 832,985 EE obs. R^2 adj. 91.1% SE adjusted for 640,214 clusters of workers Coefficient	Robust SE	t	Men 445,902 EE obs. R^2 adj. 91.2% SE adjusted for 337,629 clusters of workers Coefficient	Robust SE	t	Women 387,083 EE obs. R^2 adj. 91% SE adjusted for 302,585 clusters of workers Coefficient	Robust SE	t
	Subordinates	−0.639***	0.004	−176.4	−0.653***	0.005	−121.1	−0.632***	0.005	−128.5
	3rd officers and specialists	−0.649***	0.003	−196.9	−0.647***	0.005	−136.7	−0.649***	0.005	−137.2
	Administrative assistants	−0.653***	0.003	−211.9	−0.655***	0.005	−128.1	−0.651***	0.004	−163.7
	Over 18 years unqualified or under 18 years	−0.742***	0.003	−228.1	−0.750***	0.005	−157.1	−0.720***	0.004	−160.7
Activity sector (reference: Public administration)	Financial and business services	0.302***	0.005	58.0	0.251***	0.008	32.0	0.347***	0.007	49.1
	Supplies	0.285***	0.006	47.8	0.239***	0.008	29.9	0.329***	0.011	28.9
	Extractive and manufacturing industries	0.237***	0.004	62.2	0.205***	0.006	33.7	0.248***	0.006	44.2
	Construction	0.228***	0.004	61.0	0.201***	0.006	33.1	0.305***	0.007	43.4
	Other services	0.087***	0.004	23.1	0.069***	0.006	11.2	0.094***	0.005	19.6
	Trade, hotels & restaurants, transport & communic.	0.060***	0.004	16.6	0.040***	0.006	6.8	0.068***	0.005	14.6
	Health	−0.024***	0.004	−6.4	−0.002	0.007	−0.3	−0.027***	0.005	−5.9
	Education	−0.081***	0.003	−24.1	−0.081***	0.006	−13.5	−0.085***	0.004	−20.8
	Agriculture	−0.158***	0.005	−31.7	−0.187***	0.007	−25.9	−0.106***	0.008	−13.9
Public sector workers		0.691***	0.008	89.1	0.783***	0.011	70.9	0.6***	0.011	55.3
Constant		3.981***	0.012	343.1	3.961***	0.017	237.9	4.133***	0.016	255.4

Note(s): *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$. Additional control variables: Worker attributes (Nationality, Social benefits received in 2021, Income received in 2021 from professional activities, Number of labour contracts with the company during 2021, Duration since the first contract with the company (years), Collective agreement); Job placement attributes (Province, Type of contract, Percentage of the income from work that is in kind); Firm attributes (Legal person vs. natural person, Number of workers (logs)). The complete table is offered as supplementary material (available online)

Source(s): Authors' own work based on MCVL