

Have women really a better access to best-paid jobs in the public sector?

Counterfactuals based on a job assignment model*

Laurent Gobillon[†]

Dominique Meurs[‡]

Sébastien Roux[§]

October 12, 2015

Abstract

We propose a job assignment model with individual heterogeneity and derive a measure of gender differences in the propensity to get job positions along the wage distribution. This measure is used to quantify gender disparities in the public and private sectors with a French exhaustive administrative dataset. We then show how the model can motivate decomposition and counterfactual exercises consistent with economic theory. Our approach constitutes an alternative to more descriptive methods such as Oaxaca decompositions and quantile counterfactual approaches. Gender differences in the propensity to get jobs along the wage distribution are found to be rather similar in the public and private sectors, although the gap is slightly larger in the public sector in the 0.5 – 0.85 rank interval (and slightly smaller at ranks above 0.85). The overall contribution of observables to explaining gender differences is small in the two sectors except for ranks below 0.5 in the public sector. Long part-time experience is the only factor with a sizable explanatory power. The gender wage gap in the public sector increases by only 0.7 points when workers are assigned to jobs according to the rules of the private sector, but the gender quantile gap at the last decile is significantly larger by 4.6 points.

Keywords: gender, discrimination, wages, quantiles, job assignment model, glass ceiling, public sector.

JEL Classification: J16, J31, J71

*We thank Cecilia Garcia-Penalosa, Roland Rathelot, Antoine Terracol as well as seminar and conference participants at University of Paris 1, University of Rennes, World Congress 2015 of the Econometric Society in Montreal, AFSE 2014 in Lyon, TEPP 2015 in Aussois, *Workshop on Gender and Ethnic Differentials in Market Outcomes* at AMSE, and OFCE. All remaining errors are ours.

[†]Paris School of Economics-CNRS, INED, CEPR and IZA. Address: Paris School of Economics (PSE), 48 Boulevard Jourdan, 75014 Paris, France. Email: laurent.gobillon@psemail.eu.

[‡]University of Paris Ouest Nanterre (EconomiX) and INED. Address: Institut National d'Etudes Démographiques (INED), 133 Boulevard Davout, 75980 Paris Cedex 20, France. Email: dominique.meurs@ined.fr.

[§]Banque de France and CREST. Centre de Recherche en Economie et Statistique (CREST), 15 Boulevard Gabriel Péri, 92245 Malakoff Cedex, France. Email: sebastien.roux@ensae.fr.

1 Introduction

There is now a large body of evidence showing sizable gender wage differences related to the segregation of females in low-paid jobs (Meurs and Ponthieux, 2015). In this paper, we propose a new way to quantify the role of observable characteristics in explaining the gender wage gap based on a job assignment model where job positions are characterized by heterogeneous wages and there are gender differences in the propensity to get these positions.

In our model, job positions are indexed by their rank in the wage distribution. All workers consider that the position at the highest rank is the most attractive in terms of earnings but some of them find work conditions too constraining to apply. Observable characteristics, including gender, can influence both the propensity to apply and applicants' chances of being selected for the job. Workers not selected for the position consider the job at the second highest rank, and so on, until all job positions are filled. Overall, this model can be seen as a way of assigning workers to jobs depending on the result of careers meant to get the best paid positions. It can be brought to the data and the effects of observable characteristics on the propensity to get jobs along the wage distribution can be estimated. It is then possible to make gender wage decompositions and construct gender wage counterfactuals by fixing parameters related to the propensity to get jobs to alternative values.

As our approach is based on economic theory, it provides an alternative to more descriptive methods such as gender quantile decompositions and counterfactual distributions (Melly, 2005a; Machado and Mata, 2006; Firpo, Fortin and Lemieux, 2009; Rothe, 2012; Chernozhukov, Fernandez-Val and Melly, 2013).¹ In our setting, counterfactuals are the result of an equilibrium when modifying structural parameters. They are generated by changing the way workers are assigned to jobs and counterfactual assignments affect the wage distribution conditional on individual characteristics but not the overall wage distribution which is held fixed. By contrast, the traditional descriptive approach considers counterfactuals of conditional wage distributions computed from another sample and this generates an overall wage distribution that depends on the counterfactual. Moreover, equilibrium effects leading to a reassignment of workers across job positions are ignored. There is also an important difference in the definition of the rank when studying gender differences across ranks in the wage distribution. Whereas we consider the rank in the wage distribution of job positions which is a common index for males and females, gender quantile differences involve two different ranks corresponding to the ranks in the two gender wage distributions.

We show that it is possible to estimate a flexible semi-parametric version of the model by maximum likelihood to recover the influence of observable characteristics including gender on the propensity to get jobs. Indeed, the probability of a job being filled with a specific worker rather than the other available workers considering the job is given by a logit model with the additional feature that the coefficients of explanatory variables are gender-

¹For a survey, see Firpo, Fortin and Lemieux (2011).

specific polynomial series of the rank. We introduce interactions between gender and observable characteristics in the specification to allow for gender-specific probability of being selected to depend on observable characteristics and the rank. Counterfactuals of wage distributions are generated by changing the coefficients of explanatory variables that capture differences in access to jobs. Consistency when the number of individuals tends to infinity is proved by extending results in sampling theory proposed by Rosén (1972), as it is possible to draw a parallel between the selection of workers at each rank and the sampling of observations without replacement. We propose a simulation approach that yields consistent estimators of these counterfactuals when the number of simulations tends to infinity.

As an illustration, we apply our approach to assess the respective importance of gender differences in the public and private sectors. It is said that females may be treated more fairly in the public sector because recruitments and promotions are based on competition, and labor unions are strong. The public sector is indeed characterized by a smaller wage gap which is consistent with these arguments. However, the wage dispersion is also smaller and may hide an assignment to jobs that is not that favorable to females. We first estimate the model on each sector separately and recover structural parameters measuring the way workers are assigned to jobs in each sector depending on their gender and other observable characteristics. Our work complements the literature on gender earnings differences which was originally cross-section (Albrecht, Björklund and Vroman, 2002) but has developed toward dynamic approaches to take into account the forward-looking behaviour of workers and evaluate specific mechanisms making use of individual time variations (Bowlus, 1997; Flabbi, 2010a and 2010b; Gayle and Golan, 2011). In contrast with this literature, our goal is to propose a measure of gender differences in propensity to get positions along a job hierarchy in a context where jobs are heterogeneous using a job assignment model that is closer in the spirit to Sattinger (1975) and Teulings (1995).²

We then compare the profile of gender differences in propensity to get jobs along the wage distribution of job positions between the two sectors and assess the role of observable characteristics in explaining inter-sectorial dissimilarities. For that purpose, we estimate counterfactuals of gender wage distributions in the public sector considering that workers in that sector are assigned the same way whatever their gender, or alternatively considering that their assignment follows the rule of the private sector. Our work adds to the literature on public-private differences which has mostly used standard Oaxaca decompositions (Lucifora and Meurs, 2006) and gender quantile decompositions (Melly, 2005b), although some recent papers estimate structural dynamic models (Postel-Vinay and Turon, 2007; Bradley, Postel-Vinay and Turon, 2013).³

For our empirical application, we rely on the *DADS Grand Format - EDP 2011* which is a unique administrative panel dataset recording all jobs in the public and private sectors over the 1992-2011 period for all workers born in the first four days of October. In particular, it provides accurate information on wages and the longitudinal

²See Sattinger (1993) for a survey on early literature.

³See also Gregory and Borland (1999) for a survey on earlier literature.

dimension can be used to reconstruct the histories of part-time activity and work interruptions. Estimations are conducted for the year 2011 considering only full-time jobs for workers aged 30-65 to avoid early-career and unstable job positions.

We find that the gender probability ratio of getting a given job along the wage distribution are rather similar in the public and private sectors, although the gap is slightly larger in the public sector in the 0.5 – 0.85 rank interval (and slightly smaller above rank 0.85). In each sector, the overall contribution of observables (age, diploma, part-time history, work interruption history, and location in Paris region) to explaining gender differences in propensity to get jobs is small except for ranks below 0.5 in the public sector. Long part-time experience is the only factor that impedes females to get jobs to some extent. The raw average gender wage gap in the public sector at 13.2% is smaller than that in the private sector which stands at 15.3%. Interestingly, when workers in the public sector are assigned to jobs according to the rules of the private sector, the gender wage gap is only 13.9%. This suggests that the gender wage gap difference between the two sectors due to differences in assignment rules would be rather small, around 0.7 points, and the raw difference of 2.1 points would be mostly due to the larger wage dispersion in the private sector. By contrast, the change in gender quantile gap at the last decile when assigning workers in the public sector with the rules of the private sector is large as it reaches 4.6 points.

In Section 2, we explain how counterfactuals can be constructed from a job assignment model by changing the propensities to get jobs. We detail in Section 3 how the model can be estimated and empirical counterparts of counterfactuals can be obtained. Section 4 presents the French public sector as well as the data and provides descriptive statistics. Section 5 gives the results on estimated coefficients and counterfactuals. Finally, Section 7 concludes.

2 The model

We consider each sector separately and ground our analysis within each sector in the job assignment framework proposed by Gobillon, Meurs and Roux (2015) extended to take into account the observable heterogeneity of workers. There is an infinite but countable number of workers such that there is a proportion $n(m)$ of males in the population, which we refer to as the *measure* of males for clarity hereafter, and a proportion $n(f) = 1 - n(m)$ of females. Workers are characterized by observable attributes X which will affect their chances of getting a job. We focus on the case where attributes take a finite number of values $\{X^k\}_{k=1,\dots,K}$ with K the number of values. Denoting $n(X, j)$ the measure of gender- j workers with characteristics X and $F_{X,j}(\cdot)$ the cumulative distribution of X for the population of workers with that gender, we have $\int n(x, j) dF_{X,j}(x) = n(j)$.

Job positions are heterogenous such that they can be ranked according to a continuous variable related to the position in the job hierarchy.⁴ Consistently with our application, we consider that this variable is the wage and

⁴We assume that the variable used to rank jobs is continuous as we want to estimate the parameters of the model from wage data.

we assume that each job position is associated with a specific fixed wage through a contract. This wage is not allowed to depend on the gender or the other observable characteristics of the applicant. We suppose that two job positions cannot be associated with the same wage offer so that each job can be uniquely identified by its *rank* in the wage distribution.⁵ Workers first decide whether or not to apply for the best ranked position as it offers the highest wage but work conditions may be too constraining. The applicants are all in competition whatever their characteristics. The manager of the position chooses an applicant taking into account the observable attributes of all applicants. Those not hired turn to the second best ranked position, and so on.

More formally, the recruitment process can be described in the following way. For a job of rank u , a manager screens the available workers who decide to apply. We denote by $n(u|X, j)$ the measure of gender- j workers with characteristics X who are available for a job of rank u such that $n(1|X, j) = n(X, j)$ as all workers are available for the first job and $n(0|X, j) = 0$ as all workers end up being employed. Workers are heterogeneous in their labour supply. For instance, some females may give up applying to well-paid jobs when they have children because these jobs involve long working hours and investment not compatible with family constraints. Denoting by $v(u|X, j)$ the exogenous share of available gender- j workers with characteristics X who decide to apply for the job, the measure of applicants with such characteristics is given by $v(u|X, j) n(u|X, j)$. The manager derives a utility $V_i(u)$ from an applicant i that captures both firm profit and taste, and is supposed to take the form:

$$V_i(u) = \ln \varphi(u|X_i, j(i)) + \varepsilon_i(u) \quad (1)$$

where $\varphi(u|X, j)$ captures both the profit component that depends on observables and the taste of the manager for workers of gender j and characteristics X , and $\varepsilon_i(u)$ captures a random profit component derived from the match (ie. the match quality) drawn independently across individuals. We consider that the manager observes the two profit components and hires the applicant yielding the highest utility. The set of applicants to the job is the set of workers not hired for a job of higher rank and interested in the job. This set can be defined recursively as:

$$\Omega(u) = \left\{ i \text{ applying for rank-}u \text{ job} \mid \text{for all } \tilde{u} > u, i \text{ not applying for rank-}\tilde{u} \text{ job or } V_i(\tilde{u}) < \max_{k \in \Omega(\tilde{u})} V_k(\tilde{u}) \right\} \quad (2)$$

The set of applicants for the job, $\Omega(u)$, contains all the workers who did not apply for the jobs above rank u because of work constraints or did not draw a random profit component high enough to get selected for those jobs.

For a given job, the maximization program of the manager is a multinomial model with two specificities. First, the choice set consists in all applicants still available after better ranked job positions have been filled. To avoid any selection process based on match quality, we suppose that match qualities are drawn independently across

Nevertheless, the model can easily be adapted to the case where this variable is discrete.

⁵This assumption is made for expositional purposes and can easily be relaxed at the cost of a slightly less straightforward presentation. Indeed, formulas become a bit more intricate when several job positions are filled simultaneously because workers are interested by all of them to the same extent.

jobs. Second, the choice set contains an infinite but countable number of workers. We extend the extreme value assumption on the law of residuals that is associated with a logit specification to an infinite countable number of jobs following Dagsvik (1994).⁶ This assumption ensures that for any given job, the probability of selecting a worker follows a logit model. Under this assumption, the probability that the worker chosen for the job of rank u is of gender j and has characteristics X verifies:

$$P(j(u) = j, X(u) = X) = n(u|X, j) \phi(u|X, j) \quad (3)$$

with:

$$\phi(u|X, j) = \frac{\mu(u|X, j)}{\int n(u|X, f) \mu(u|X, f) dX + \int n(u|X, m) \mu(u|X, m) dX} \quad (4)$$

where $\mu(u|X, j) = v(u|X, j) \varphi(u|X, j)$ captures labour supply, productivity and taste effects.⁷ In our application, we will not try to disentangle these three types of effects. We will rather evaluate the overall effects of explanatory variables on the propensity to get a job of rank u for a gender- j worker with characteristics X , $\mu(u|X, j)$, that we label from now on the “conditional worker value”. The denominator of the right-hand side term in (4) is a competition term which depends on the measures of applicants with the different genders and characteristics. We will refer to $\phi(u|X, j)$ as the conditional probability of getting the job of rank u , and to $\phi(u|X, m) / \phi(u|X, f) = \mu(u|X, m) / \mu(u|X, f)$ as the gender conditional probability ratio of getting the job for workers with characteristics X .⁸

For gender j , we can derive a differential equation verified by the measure of individuals with characteristics X available for a job of rank u . Consider an arbitrarily small interval du in the unit interval. The proportion of jobs in this small interval is du since ranks are equally spaced (and dense) in the unit interval. The measure of jobs occupied by workers of a given gender j with characteristics X is $n(u|X, j) \phi(u|X, j) du$. For these gender and characteristics, the measure of workers available for a job of rank $u - du$ can be deduced from the measure of workers available for a job of rank u subtracting the workers who get the jobs of ranks between $u - du$ and u :

$$n(u - du|X, j) = n(u|X, j) - n(u|X, j) \phi(u|X, j) du \quad (5)$$

⁶Put differently, we assume that the points of the sequence $\{j(i), X_i, \varepsilon_i(u)\}$, $i \in \Omega(u)$ are the points of a Poisson process with intensity measure $\frac{n(u|X, j) \mu(u|X, j)}{\int n(u|X, f) \mu(u|X, f) dX + \int n(u|X, m) \mu(u|X, m) dX} \exp(-\varepsilon) d\varepsilon$.

⁷Our framework can also be extended to take into account statistical discrimination with $\mu(u|X, j)$ capturing its intensity on top of the other mechanisms (see Gobillon, Meurs and Roux, 2015).

⁸Note that the probability of getting the job at a given rank $\phi(u|X, j)$ is conditional not only on the specific characteristics of a given applicant, but also on the characteristics of all workers not hired for a job of higher rank and interested in the job. We do not write explicitly this second conditioning to keep notations simple.

From this equation, we obtain when $du \rightarrow 0$:

$$n'(u|X, j) = \phi(u|X, j) n(u|X, j) \quad (6)$$

This relationship states that the variations in the measure of gender- j workers with characteristics X around rank u depend on the stock of gender- j workers with these characteristics and their chances of getting a job. We also show in Appendix A that, under the initial conditions $n(1|X, j) = n(X, j)$, the system of equations considering (6) for all X and j , where $\phi(u|X, j)$ verifies (4), has a unique solution.

2.1 Decompositions

Suppose that we are able to construct estimators of conditional worker values at any given rank, $\mu(u|X, j)$. It is possible to make a decomposition of the gender conditional probability ratio of getting a job into the contribution of the differences in observable characteristics and the contribution of the differences in their returns. Indeed, denote by $\phi(u|j)$ the probability of an available gender- j worker getting a job at rank u . This probability verifies:

$$\phi(u|j) = \int p(u|X, j) \phi(u|X, j) dX \quad (7)$$

with $p(u|X, j) = n(u|X, j) / n(u|j)$ where $n(u|j) = \int n(u|X, j) dX$ is the proportion of gender- j workers with characteristics X still available for a job at rank u . The (unconditional) gender probability ratio of getting a job of rank u is given by $\phi(u|f) / \phi(u|m)$. We introduce conditional worker values considered to be the benchmark in absence of any gender difference in propensity to get jobs, $\mu^r(u|X)$, which are fixed or estimable. For instance, they can be the conditional worker values for males or the conditional worker value for the overall population (in line with Oaxaca and Ransom, 1994). Taking the logarithm of the gender probability ratio of getting the job of rank u derived from (7) and rearranging the terms, we get:

$$\begin{aligned} \log [\phi(u|f) / \phi(u|m)] &= \int [p(u|X, f) - p(u|X, m)] \log \mu^r(u|X) dX \\ &+ \int [\log \mu(u|X, f) - \log \mu^r(u|X)] p(u|X, f) dX \\ &- \int [\log \mu(u|X, m) - \log \mu^r(u|X)] p(u|X, m) dX + r(u) \end{aligned} \quad (8)$$

where:

$$\begin{aligned} r(u) &= \left[\int p(u|X, f) \log \mu(u|X, f) dX - \log \left[\int p(u|X, f) \mu(u|X, f) dX \right] \right] \\ &- \left[\int p(u|X, m) \log \mu(u|X, m) dX - \log \left[\int p(u|X, m) \mu(u|X, m) dX \right] \right] \end{aligned} \quad (9)$$

The first right-hand side term in (8) reflects the gender difference in propensity to get the job at a given rank for available workers if conditional worker values are the same for males and females, and fixed to the benchmark values. This gender difference is due only to gender differences in the composition of available workers. The second (resp. third) right-hand side term reflects the gender difference in propensity to get the job if conditional values of available female (resp. male) workers were modified to take the benchmark values. The fourth one is the residual due to the non-linearity introduced by the use of logarithms. All right-hand side terms can be computed replacing conditional worker values by their estimators.

Importantly, the set of workers available at each rank is fixed and determined from the data. We do not reassign workers to jobs when changing conditional worker values. We now show how to perform counterfactual exercises that involve a reassignment of workers when changing conditional worker values.

2.2 Counterfactuals

A matter of interest is the gender differences in propensity to get each job along the wage distribution in the public sector if individuals were attributed the conditional worker values in the private sector, or if individuals of the same gender were attributed the same conditional worker values. To investigate this matter, we construct counterfactuals of gender conditional probabilities of getting jobs if gender- j workers were attributed some alternative conditional worker values $\mu^*(u|X, j)$. In this counterfactual situation, there is a reassignment of workers across job positions holding fixed the wage distribution of job positions. Denote by $n^*(u|X, j)$ the counterfactual assignment of workers which is obtained from the differential equation (6) where the conditional probability of getting a job has been replaced by its expression (4) and conditional worker values by their counterfactuals:

$$n^{*'}(u|X, j) = \frac{\mu^*(u|X, j) n^*(u|X, j)}{\sum_{\ell, g} \mu^*(u|X^\ell, g) n^*(u|X^\ell, g)} \quad (10)$$

This differential equation is solved under the initial conditions $n^*(1|X, j) = n(1|X, j)$ where $n(1|X, j)$ is the measure of gender- j workers with characteristics X in the public sector.

The counterfactual of the gender probability ratio of getting a job can easily be obtained by replacing the conditional probabilities of getting this job by their expressions (7) where conditional worker values have been replaced by their counterfactuals and the proportion of gender- j workers with characteristics X still available for a job at rank u by the counterfactual $p^*(u|X, j) = n^*(u|X, j)/n^*(u|j)$ with $n^*(u|j) = \int n^*(u|X, j) dX$. The counterfactual of the gender probability of getting the job is given by:

$$\phi^*(u|j) = \int p^*(u|X, j) \mu^*(u|X, j) dX \quad (11)$$

We are also interested in constructing gender-specific counterfactual wage distributions and, in particular, gender-

specific wage distributions for workers in the public sector if the conditional values of these workers were modified to be the same as in the private sector. The differences between the actual gender-specific wage distributions and their counterfactuals are due to differences in propensity to get jobs between the two sectors.

The counterfactual of gender- j wage cumulative in the public sector when workers have the same conditional values as in the private sector verifies:

$$F_j^*(w) = \int \frac{n^*(F(w)|X, j)}{n^*(1|j)} dX \quad (12)$$

where $F(\cdot)$ is the wage cumulative in the public sector and $n^*(1|j) = \int n^*(1|X, j) dX$ is the counterfactual measure of gender- j workers. The derivation of this relationship gives the counterfactual of gender- j wage density:

$$f_j^*(w) = f(w) \int \frac{n^{*'}(F(w)|X, j)}{n^*(1|j)} dX \quad (13)$$

where $f(\cdot)$ is the wage density in the public sector.

3 Empirical strategy

3.1 Estimation of parameters

It is possible to quantify the influence of observable characteristics on conditional worker values under semi-parametric assumptions. We make the assumption that conditional worker values can be specified as:

$$\mu(u|X, j) = \exp[X\beta_j(u)] \quad (14)$$

where X now refers to a matrix of attributes which includes the vector of ones and $\beta_j(\cdot)$ are some gender-specific functions of the rank that we choose to be polynomials of finite order. This model makes an index assumption to decrease the dimensionality but coefficients that can depend on the rank. This dependence is allowed because labour supply may vary across rank in a specific way for workers depending on their characteristics, or because characteristics may be valued differently by employers depending on the position that is considered. In that setting, the empirical counterpart of the conditional probability of a worker i getting a job at rank u given by (4) is simply a logit model such that the latent variable associated to the worker is $X_i\beta_{j(i)}(u) + \eta_i(u)$ with $\eta_i(u)$ following an extreme value law.⁹ This latent variable looks like the utility derived by the manager from the worker being hired (1) except that the coefficients $\beta_j(u)$ do not measure the effects of explanatory variables on the manager's tastes only, but rather they joint effects on these tastes, productivity and labour supply.

⁹Note that the residual $\eta_i(u)$ is not equal to the productivity match $\varepsilon_i(u)$ because of heterogeneous labor supply effects. We have $\eta_i(u) = \varepsilon_i(u)$ when supply effects are uniform across all individuals.

The parameters of polynomial coefficients can be estimated by maximum likelihood. We first introduce some additional notations. Denote by u_i the rank of individual i and X_i the value of his observable attributes, $u^k = (k-1)/(N-1)$ the k^{th} rank, i_k the individual occupying this rank such that we have $\Omega(u^k) = \{i_1, \dots, i_k\}$, and $\vec{X}_k = \{X_{i_1}, \dots, X_{i_k}\}$ and $\vec{j}_k = \{j(i_1), \dots, j(i_k)\}$ respectively the observed characteristics and genders of the individuals occupying the k lowest paid jobs. The likelihood is given by:

$$\begin{aligned}
L &= P\left(u_{i_1} = u^1, u_{i_2} = u^2, \dots, u_{i_N} = u^N \mid \vec{X}_N, \vec{j}_N\right) \\
&= P\left(u_{i_N} = u^N \mid \vec{X}_N, \vec{j}_N\right) \prod_{k=1}^{N-1} P\left(u_{i_k} = u^k \mid u_{i_{k+1}} = u^{k+1}, \dots, u_{i_N} = u^N, \vec{X}_N, \vec{j}_N\right) \\
&= P\left(u_{i_N} = u^N \mid \vec{X}_N, \vec{j}_N\right) \prod_{k=1}^{N-1} P\left(u_{i_k} = u^k \mid u_{i_{k+1}} = u^{k+1}, \dots, u_{i_N} = u^N, \Omega(u^k), \vec{X}_k, \vec{j}_k\right) \\
&= \prod_{k=1}^N P\left(u_{i_k} = u^k \mid \Omega(u^k), \vec{X}_k, \vec{j}_k\right)
\end{aligned}$$

where the last equality is obtained under our assumption that the random match qualities $\varepsilon_i(u)$ are independently and identically distributed across ranks. Here, $P\left(u_{i_k} = u^k \mid \Omega(u^k), \vec{X}_k, \vec{j}_k\right)$ is the empirical counterpart of $\phi(u_{i_k} \mid X_{i_k}, j(i_k))$ and it verifies:

$$P\left(u_{i_k} = u^k \mid \Omega(u^k), \vec{X}_k, \vec{j}_k\right) = \frac{\mu(u_{i_k} \mid X_{i_k}, j(i_k))}{\sum_{\ell \leq k} \mu(u_{i_\ell} \mid X_{i_\ell}, j(i_\ell))} = \frac{\exp\left[X_{i_k} \beta_{j(i_k)}(u_{i_k})\right]}{\sum_{\ell \leq k} \exp\left[X_{i_\ell} \beta_{j(i_\ell)}(u_{i_\ell})\right]} \quad (15)$$

The parameters of polynomial coefficients $\beta_j(u)$ are estimated by maximizing the log likelihood $L = \frac{1}{N} \sum_i P\left(u_{i_k} = u^k \mid \Omega(u^k), \vec{X}_k, \vec{j}_k\right)$. In fact, the likelihood is the same as the partial likelihood obtained when estimating a Cox duration model with time-varying parameters, and the asymptotic distribution of estimated parameters has long been established (see Andersen and Gill, 1982).

3.2 Evaluation of decompositions and counterfactuals

It is also possible to make an empirical assessment of the decomposition of the gender probability ratio of getting each job given by (8). Assume that benchmark conditional worker values $\mu^r(u \mid X)$ are also of the form (14) with estimable polynomial coefficients $\beta^r(u)$. In that setting, the decomposition simplifies to:

$$\begin{aligned}
\log[\phi(u \mid f) / \phi(u \mid m)] &= [E(X \mid f, u) - E(X \mid m, u)] \beta^r(u) \\
&\quad + E(X \mid f, u) [\beta_f(u) - \beta^r(u)] - E(X \mid m, u) [\beta_m(u) - \beta^r(u)] + r(u) \quad (16)
\end{aligned}$$

where $E(X|j, u)$ are the average characteristics of gender- j workers available for a job at rank u . The left-hand side term, the gender probability ratio of getting a job of rank u , can be estimated non parametrically following Gobillon, Meurs and Roux (2015). Right-hand side terms can be obtained by replacing average characteristics by their empirical counterparts, and polynomial coefficients by their estimators. When the benchmark values are those of males, $\beta^r(u) \equiv \beta_m(u)$ which have already been estimated. If the benchmark values are average conditional worker values for the overall population, $\beta^r(u)$ can be obtained by maximum likelihood fixing $\beta_f(u) = \beta_m(u) \equiv \beta^r(u)$ and adding a gender dummy to the specification to act as a control in line with Fortin (2008).

It is possible to estimate counterfactuals of gender- j probabilities of getting each job given by (11). For that purpose, we need to recover estimators of $n^*(u|X, j)$ when using the counterfactual conditional worker values $\mu^*(u|X, j)$ which we suppose to be of the form (14) with estimable polynomial coefficients $\beta_j^*(u)$. The finite discrete counterpart of the differential equation verified by the measures of available workers (6) is:

$$N^*(u^k|X, j) = N^*(u^{k+1}|X, j) - D_{k+1}(X, j) \quad (17)$$

where $N^*(u|X, j)$ is the counterfactual number of gender- j workers with characteristics X available at rank u , and $D_k(X, j)$ is a dummy taking the value one if an available worker in the set $\Omega_j^*(u^k, X)$ gets the job at rank u^k and zero otherwise. There is some randomness which comes from the choice of a worker by the manager from conditional probabilities of getting jobs of the form (4). For a given rank v , a quantity of interest is $E[N^*(u^{\lfloor vN \rfloor + 1}|X, j)]$ where $\lfloor \cdot \rfloor$ is the integer part. Indeed, we show in Appendix B that $E[N^*(u^{\lfloor vN \rfloor + 1}|X, j)]/N \xrightarrow{a.s.} n^*(v|X, j)$ for all $v \in]0, 1[$. The proof relies on the extension of a theorem on sampling without replacement proposed by Rosén (1972). Whereas in the original theorem, the influence of explanatory variables on the propensity of an individual to get the job does not vary across jobs, in our case this propensity varies since $\exp[X\beta_j(u)]$ depends on the rank. We show that the proof of the original theorem can be generalized to the case where the influence of explanatory variables varies across jobs.

The expectation $E[N^*(u^{\lfloor vN \rfloor + 1}|X, j)/N]$ can be estimated using a simulation approach which consists in generating assignments of workers to jobs. Theoretical foundations of the simulation approach are detailed in Appendix C. A simulation is indexed by $s = 1, \dots, S$, we denote by $\Omega_j^s(u, X)$ the counterfactual sample of gender- j workers with characteristics X available at rank u . In practice, the counterfactual sets of available workers with every characteristics X at the empirical rank u^{k-1} , $\{\Omega_j^s(u^{k-1}, X)\}_{X, j}$, are deduced from the same sets at next empirical rank, $\{\Omega_j^s(u^k, X)\}_{X, j}$, by subtracting the worker that gets the job at rank u^k which is determined consistently with the model specification in the following way.

The empirical counterpart of the counterfactual conditional probability of getting a job at rank u^k can be

rewritten as:

$$P\left(u_i = u^k \mid \Omega^s(u^k), \vec{X}_k^s, \vec{j}_k^s\right) = \frac{\exp\left[X_i \beta_{j(i)}^*(u^k)\right]}{\sum_{k \in \Omega^s(u^k)} \exp\left[X_k \beta_{j(k)}^*(u^k)\right]} \quad (18)$$

with $j(i)$ the gender of worker i , $\Omega^s(u^k)$ the counterfactual sample of workers available at rank u^k , and $(\vec{X}_k^s, \vec{j}_k^s)$ their characteristics and genders. It thus corresponds to the probability of getting the job given by a multinomial logit model. Hence, it is possible to simulate which worker gets the job at rank u^k by computing, for each available individual, the sum $X_i \widehat{\beta}_{j(i)}^*(u^k) + \eta_i^s(u^k)$ where $\eta_i^s(u^k)$ has been drawn in an extreme value law. The worker getting the job is the one with the highest value of this sum. The sets of available workers with given characteristics and gender at each empirical rank is obtained by recursively applying this procedure from the highest rank to the lowest rank. Finally, we deduce the number of available workers $N^s(u^k | X, j) = \text{Card } \Omega_j^s(u^k, X)$. We show in Appendix C that when $S \rightarrow +\infty$, we have for all $v \in]0, 1[$, $\frac{1}{S} \sum_{s=1}^S N^s(u^{\lfloor vN \rfloor + 1} | X, j) \xrightarrow{a.s.} E[N^*(v | X, j)]$.

We now explain how to evaluate the probability of getting a job at each rank u^k for each gender. The empirical counterparts of the terms $p^*(u^k | X, j)$ which enter the counterfactual probability (11) are:

$$\widehat{p}^s(u^k | X, j) = \frac{N^s(u^k | X, j)}{\sum_{\ell} N^s(u^k | X^{\ell}, j)} \quad (19)$$

An estimator of the counterfactual gender- j probability of getting a job at rank u^k is then given by:

$$\widehat{\phi}^*(u^k | j) = \frac{1}{S} \sum_{s=1}^S \sum_{\ell} \widehat{p}^s(u^k | X^{\ell}, j) \exp\left[X^{\ell} \widehat{\beta}_j^*(u^k)\right] \quad (20)$$

where $\widehat{\beta}_j^*(u)$ is an estimator of $\beta_j^*(u)$.

We also want to recover the counterfactuals of gender- j cumulative and density. For that purpose, we first consider the original sample and sort wages in ascending order, denoting by w^k the k^{th} wage. It is possible to construct an estimator of the counterfactual of gender- j cumulative at rank u^k , $F_j^*(w^k)$, from equation (12) as:

$$\widehat{F}_j^*(w^k) = \frac{1}{\sum_{\ell} N(X^{\ell}, j)} \frac{1}{S} \sum_{s=1}^S N^s(F_j^*(w^k) | X^{\ell}, j) \quad (21)$$

A counterfactual of gender- j density is obtained by smoothing and deriving the counterfactual cumulative.

4 Context and descriptive statistics

4.1 The public sector in France

The French public sector accounts for around 20% of total salaried employment. Most public employees are females (61%) whereas this is not the case in the private sector (44%) (Dorothee, Le Faller and Treppoz, 2013). The public sector is divided into three subsectors: central administration which includes education (44% of employment in the public sector), local government (35%) and public health (21%). The share of local government in employment has increased over the last 10 years in line with the decentralization process occurring during that period (Dorothee and Baradji, 2014).

In France, the public sector has a highly centralized pay setting compared to the private sector. A common pay scale is applied to all subsectors, which means that the nominal value of the basic wage is the same at any given grade through the entire public sector. However, individual differences in earnings may arise from bonuses which are mostly related to the type of occupation. Due to budget constraints, the basic wage is constant in nominal terms since 2010, and there has been no major change in pay scale by occupation in the past decade. As a consequence, advancement along the pay grid is currently the main way to get a pay rise.

The French public sector is very close to the model of internal labor market proposed by Doeringer and Piore (1985). The main recruitment process is a competitive exam with diploma requirements specific to the type of occupation. There are as many competitive exams as types of occupations (policemen, judges, teachers, nurses, clerks, academics, etc.), the most prestigious one leading to careers in top management in the public administration through ENA (*Ecole Nationale d'Administration*). Once recruited, civil servants start their career at the bottom of the pay scale specific to their occupation. Mobility between the public and private sectors is very limited and occurs mainly at early stage of careers (Daussin-Benichou et al., 2014). Wage increases depend much on seniority despite recent reforms which aim at taking into account individual performance and/or local work conditions. Promotions are seldom events which also depend on seniority to some extent and are most often obtained through competitive or professional exams.

4.2 The data

Estimations are conducted on the *DADS Grand Format - EDP 2011* which is a panel dataset following all individuals born in the first four days of October and is constructed from two difference sources (*Déclaration Annuelles des Données Sociales* i.e. *DADS* and *Echantillon Démographique Permanent*, i.e. *EDP*). The data record all jobs in the public and private sectors since 1976.¹⁰

The DADS are collected for tax purposes and contain details on job characteristics. They give the establishment

¹⁰Three years, 1981, 1983 and 1990, are missing because the French Institute of Statistics devoted its financial means to the 1982 and 1990 census rather than to the management of the wage data.

identifier (*SIREN* number) from which we determine establishment seniority since 1992 and the status (full-time or part-time) from which we reconstitute part-time history. We also compute the number of years individuals are absent from the data. Absence corresponds to an interruption in the professional activity due to unemployment or exit from the labour force. The full-time equivalent annual wage is reported. There are outliers with wage below the minimum wage and consequently we delete observations for which the monthly wage is below 1000 euros. The job duration (in days) is reported and we only retain jobs occupied full time on July 1st in which workers stayed for at least 30 days during the year to focus on stable workers that are more likely to compete for all job positions. This also means that we keep at most one job per worker every year. Finally, we use the information on administration for public jobs to restrict the sample to those in the central and local administration but excluding teaching jobs which management is very specific.¹¹ We consider jobs only for workers aged 30-65 to avoid taking into account the frequent transitions between unstable job positions that often occur at the beginning of the career. We limit our attention to the single year 2011 as our assignment model is cross-section. Our final dataset contains 55,881 observations and the proportion of females is 37.8%. Most individuals work in the private sector (82.6%) where the proportion of females (35.1%) is lower than in the public sector (50.9%).

Data are also used to construct our set of explanatory variables. We consider a dummy for establishment tenure being larger or equal to 10 years and two dummies for part-time experience being respectively between 7% (the median) and 18% (the third quartile), and more than 18%, of job experience (less than 7% being the reference). The location of job at the municipality level is used to construct a dummy for the job being located in the Paris region. We also consider two dummies for the age brackets 41-50 and 50-65 (31-40 being the reference). These variables are complemented with information on diploma and children. We construct a dummy for having a high-school diploma or more, as well as two dummies for having respectively no child, and three children and more (one or two children being the reference). Note that we are rather parcimonious in the number of categories. This is because we need to estimate the gender-specific coefficients of polynomial function of ranks for each dummy in the empirical application and this makes the number of coefficients increase fast with the number of explanatory variables.

4.3 Descriptive statistics

We report descriptive statistics on wages by sector and gender in Table 1. As usual, females have a lower average wage than males. When pooling the two sectors, the gap is 19%. There is also a smaller wage dispersion for females as the standard deviation is 37% lower than that of males. The gender quantile difference increases with the rank in line with the literature, the difference being only 8% at the 5th percentile but as high as 26% at the 95th percentile. There are interesting differences between the two sectors. The public sector is characterized by less dispersed wages, a gender wage gap (14%) which is lower than that of the private sector (19%), and a gender

¹¹Note that jobs in the health administration are also excluded because some positions such as doctors are very specific and workers occupying them usually do not change job for a position in another administration of the public sector.

quantile difference which increases more slowly with the rank. These differences between the two sectors are closely related to differences in their wage dispersion.

[*Insert Table 1*]

We represent the gender log-wage distributions by sector. Figure 1 shows that gender distributions in the private sector have fatter right-hand tails and lower peaks than in the public sector. In the two sectors, male distribution is slightly to the right of female distribution, especially in the public sector.

[*Insert Figure 1*]

We also investigate gender differences in observable characteristics within and across sectors. Table 2 shows that in each sector females are more qualified than males, and have much more often long part-time experience and long work interruption. The gender gap in part-time experience is very large and is similar across sectors. By contrast, the gender gaps in education and work interruption are smaller in size but remain sizable, and they are larger in the public sector.

[*Insert Table 2*]

5 Results

5.1 Estimation of the gender probability ratio of getting a job

We compute a non-parametric estimator of the gender probability ratio of getting each job for each sector following the procedure proposed by Gobillon, Meurs and Roux (2015). Figure 2 shows that above rank 0.05, females have a lower propensity than males to get any job position whatever the sector. In the private sector, the female propensity to get jobs decreases slowly until rank 0.45 and then increases before decreasing again after rank 0.65. Non-monotonic movements can be explained by the heterogeneity of jobs as very heterogeneous industries are pooled. At the highest quantiles, female propensity to get jobs is very low: a female has 70% less chances of getting a job than a male. In the public sector, the gender probability ratio of getting jobs is decreasing until rank 0.4, nearly flat for ranks in the 0.4 – 0.9 interval, and then again decreasing. Interestingly, between rank 0.5 and 0.85, female propensity to get jobs is lower in the public sector than in the private sector. Nevertheless, it ends up being slightly higher at the highest ranks.

[*Insert Figure 2*]

We then estimate the semi-parametric model (14) when introducing all our explanatory variables, including the gender dummy but without considering its interaction with other variables, and fixing the degree of the rank of

polynomial coefficients to five for each variable. Estimated coefficients (that of gender dummy excluded) will be used as references in our counterfactual exercise in the scenario where conditional worker values for the two genders are equalized.

For this simple specification, we can compare the (exponentiated) effects of dummy explanatory variables on the gender conditional probability ratio of getting a given job between the two sectors. These effects are represented as a function of the rank in Figure 3. The profiles for the gender dummy are consistent with the non-parametric estimators of the gender probability ratio of getting each job. Controlling for observable characteristics thus does not affect much the estimated gender probability ratio of getting a job.¹² The effect of high-school diploma increases more slowly at high ranks in the public sector but exhibits a sharper increase at highest rank than in the private sector. Not surprisingly, workers with no high-school diploma have nearly no chance of getting the best-paid jobs in the two sectors. Workers with three children and more have a higher probability of getting best-paid jobs in the two sectors, especially the private one. It is possible to check that this occurs mostly because of males. It is consistent with males with three children being more stable or being ready to work when they are the main provider of the household. Job seniority is not necessary to get best-paid jobs in the private sector but it plays a large role in the public one. This can be explained by job mobility in the private sector that can lead to the occupation of better-paid positions. Age profiles are similar in the two sectors such that it is easier to get best-paid jobs when you get older. The probability of getting best-paid jobs is also larger when living in the Paris region in the two sectors, consistently with the location of best-paid jobs. Short part-time experience is mostly detrimental in the public sector. This could be due to labor supply effects such that part-time workers in that sector are less career-oriented, or to career rules as part-time entry jobs impede sectoral seniority which is important for promotions in the public sector. By contrast, long part-time experience is detrimental in the two sectors. Finally, the picture is similar when considering short/medium and long work interruptions.

[*Insert Figure 3*]

We also consider a more complete semi-parametrization of conditional worker values in which we additionally include interactions between all explanatory variables and the gender dummy. The degree of polynomial coefficients including those of interactions is again fixed to five. Figure 4 shows for each sector that the non-parametric gender probability ratio of getting a given job is in the confidence interval of the average simulated one for 100 replications when using the semi-parametric version of the model. In fact, the two curves are nearly confounded, which suggests

¹²We also tried to determine to what extent the shape of the gender probability ratio of getting a job might be influenced by unobserved individual heterogeneity. For that purpose, we added an unobserved individual heterogeneity term to the value of each worker and assessed whether this had an influence on our non-parametric estimator of the gender probability ratio of getting a job. Unobserved individual heterogeneity terms are drawn identically and independently in a centered normal law with variance $k^2.V\left(X_i\bar{\beta}_{j(i)}\right)$, where $\bar{\beta}_j = \int_0^1 \beta_j(u) du$ and k is a parameter that is made to vary to change the scale of the variance. Figure A2 shows that increasing k leads to a decrease in the slope of the gender probability ratio of getting a job both in the public and private sectors. Nevertheless, k must be large (above 1) to cause a significant decrease in the slope.

that the semi-parametric approach is rather reliable. For each variable, Figure 5 represents the exponentiated gender difference in the estimated coefficient of the variable which corresponds to the conditional gender probability ratio of getting a given job (while holding other variables to their reference values). Except for a very few exceptions, exponentiated gender differences in estimated coefficients are well below one and take lower values at higher ranks. This suggests that females have a lower propensity to get jobs than males whatever the observable characteristics, and the gender difference is larger for high-paid jobs than for low-paid jobs.

[*Insert Figures 4 and 5*]

5.2 Decompositions and counterfactuals

We then implement the Oaxaca decomposition of the gender probability ratio of getting a given job in each sector given by equation (9). Figure 6 shows that the explained part of this gender probability ratio is small at nearly all ranks in the private sector. By contrast, the explained part is rather large for ranks below 0.5 in the public sector before becoming small at higher ranks.

[*Insert Figure 6*]

It is also possible to decompose the explained part of the gender probability ratio of getting a given job into the contributions of every variable. Figure 7 suggests that it is mostly long part-time experience that has some explanatory power. In the two sectors, long part-time experience significantly decreases female propensity to get jobs along the wage distribution of job positions. In the private sector, its contribution represents around 30% of the gender probability ratio of getting a given job up to rank 0.8. In the public sector, it explains all the gender probability ratio up to rank 0.2, but its importance then decreases with the rank and becomes small after rank 0.6. High-school diploma has a small explanatory power and increases a bit female propensity to get jobs above rank 0.5 in the private sector. We then decompose the unexplained part of gender probability ratio of getting a given job into the contribution of each variable. Figure A2 shows that the contribution of every variable is negligible or very small. Overall, the Oaxaca decomposition suggests that gender differences in propensity to get jobs cannot be explained by composition effects or differences in the returns of observables.

[*Insert Figure 7*]

Counterfactual gender probability ratios of getting any given job are computed in the two sectors when the two genders are given the same conditional worker values which are fixed to their common reference. They capture gender differences in chances of getting jobs that are only related to gender differences in observable characteristics. A major difference between this approach and the Oaxaca decomposition is that now workers are re-assigned to job positions. Interestingly, results represented in Figure 8 show that there are still gender differences in the counterfactual

situation. In the public sector, female propensity to get jobs is around 18% lower than that of males for ranks below 0.8. Descriptive statistics in Table 3 and log-wage distributions in Figure 10 show that the counterfactual gender wage gap does not disappear as the gender average wage gap stands at $100 * [1 - \exp(-0.0230)] = 2.3\%$. These results suggest that it is important to take into account the equilibrium effects related to a reallocation of workers across jobs when considering counterfactuals. By contrast, the pattern differs in the private sector as the counterfactual female propensity to get jobs is only around 10% lower than that of males below rank 0.6, but it becomes slightly higher above rank 0.8. The counterfactual gender average wage gap is now closer to zero at 0.7%.

[*Insert Figure 8*]

[*Insert Table 3*]

We then consider the counterfactual situation where workers in a given sector are attributed the conditional values in the other sector. First focusing on the public sector, we can assess with this counterfactual whether females in that sector would have a worse access to jobs and lower wages if they were assigned to jobs in a similar way as workers in the private sector. Figure 9 shows that the counterfactual gender probability ratio of getting a given job in the public sector is larger for ranks in the 0.5 – 0.85 interval but lower for ranks over 0.85 than in the initial situation. In fact, it has a shape close to the one observed in the private sector.

The counterfactual gender average wage gap in the public sector reaches $100 * [1 - \exp(-0.1494)] = 13.9\%$ and is only slightly higher than the raw wage gap which is 13.2%. It contrasts with the raw wage gap in the private sector which stands at 15.3%. Differences in assignment rules between the two sectors thus only explain 0.7 points of the gender wage gap difference which stands to 2.1 points. The rest of the gender wage gap difference between the two sectors can be explained by gender differences in observable characteristics and the larger wage dispersion in the private sector. Interestingly, there are also changes in the gender quantile gap that differs across ranks when using the counterfactual assignment rule. Whereas the gender median wage gap in the public sector is 1.1 points lower in the counterfactual situation, the gender wage gap at the last quartile and the last decile are 0.3 and 4.6 points higher respectively. Overall, the gender gap in log-wage dispersion is also higher as the gender difference in standard deviation increases by 180% (ie. by 2.2 points).

Conversely, the counterfactual gender probability ratio of getting a given job in the private sector is lower in the 0.5 – 0.85 interval but higher for ranks over 0.85 than in the initial situation. Overall, its profile across ranks is similar to that of the public sector. The counterfactual gender average wage gap at 14.4% is lower than the original gap by 0.9 points. There are also changes in the gender quantile gap that differs across ranks when changing the assignment rule, but variations are not exactly at the opposite of those in the public sector. Indeed, the gender median wage gap increases by 2.0 points, whereas the gender wage gap at the last quartile increases by 0.05 points and that at the last decile decreases by 6.6 points. The gender difference in standard deviation decreases by 65%

(i.e. by 2.2 points).

[*Insert Figures 9 and 10*]

5.3 Robustness check: hourly wage and part-time work

So far, we have conducted our analysis on the subsample of full-time workers, but this can lead to biases in the estimates if there are selection effects. It is possible that in a given sector a larger share of females ends up in part-time jobs and ignoring them leads to an underestimation of the gender differences in the propensity to get full-time job positions. Moreover, we use the daily wage to rank job positions whereas there can be gender differences in the number of hours worked. If females work on average less hours in some positions than males, their daily wage is likely to be lower and the rank of the position they hold in the job hierarchy is underestimated. Not taking into account the number of hours worked leads to an overestimation of the gender differences in the propensity to get job positions at some ranks, and an underestimation of this gender difference at ranks below them.

When repeating our analysis on a sample including part-time workers in addition to full-time ones while using the hourly wage instead of the daily wage, we obtain results that are qualitatively similar although quantitatively slightly different from those in our benchmark case for the counterfactuals in the different scenarios.¹³ For instance, the gender wage gaps at the mean and at the last decile in the public sector when using the assignment rules in the private sector are now respectively 1.4 points and 3.5 points higher than when considering the assignment rules in the public sector (instead of 0.7 and 4.6 points respectively in the benchmark case).

6 Conclusion

In this paper, we have proposed a new way to quantify the gender wage gap which is based on a job assignment model involving heterogeneous workers and job positions, as well as gender differences in the propensity to get these positions. Workers differ in their observable characteristics and they are primarily interested in the best-paid jobs. Some workers do not apply though because they consider that work constraints are too hard. Workers not getting best-paid jobs consider jobs that are paid slightly less, and so on, until all job positions are filled. Our model can be estimated using a flexible semi-parametric approach. It is also possible to construct counterfactuals of gender wage distributions when changing workers' propensity to get jobs conditional on their observable characteristics. Equilibrium effects characterized by the reallocation of workers across jobs are taken into account in these counterfactuals.

As an illustration, we apply our approach to the study of gender wage differences in the public and private

¹³Results are also qualitatively similar when we consider a sample including both the full-time and part-time workers but we use the daily wage, and when we consider a sample of full-time workers only but we use the hourly wage.

sectors for full-time workers using an original administrative dataset with accurate wage information. The gender gap in propensity to get jobs along the wage distribution is found to be rather similar in the two sectors. Overall, it has a mostly decreasing profile with values close to zero at lowest ranks but as high as 60% at highest ranks. This gender gap can hardly be explained by composition effects and gender differences in the influence of observable characteristics. It is interesting to note though that long part-time experience is significantly associated to a lower female propensity to get jobs at low ranks. Results of a counterfactual exercise show that the gender average wage gap would be slightly higher in the public sector if workers were attributed the same propensity to get jobs as in the private sector conditionally on their observable characteristics.

Several extensions of this work are worthy of interest. First, workers' choice between the two sectors could be modelled. Second, the analysis could involve a wage setting that depends on workers' observable characteristics. Finally, the framework is static and it could be interesting to design a dynamic model similar in spirit that incorporates both intra-firm and inter-firm mobilities.

7 Appendices

7.1 Appendix A: solution of the model

We have the following theorem showing the existence and uniqueness of the solution:

Theorem 1 *Suppose that X can only take a finite number of values X^ℓ , $\ell = 1, \dots, L$; $\mu(\cdot | X^\ell, m)$ and $\mu(\cdot | X^\ell, j)$ are C^1 on $(0, 1]$ for each ℓ ; and there is a constant $c > 0$ such that $\mu(u | X^\ell, m) > c$ and $\mu(u | X^\ell, f) > c$ for all $u \in (0, 1]$ and all ℓ . Then there is a unique $2L$ -uplet $\{n(\cdot | X^\ell, f), n(\cdot | X^\ell, m)\}_{\ell=1, \dots, L}$ verifying (6) where $\phi(u | X, j)$ is given by (4).*

Proof. The proof revolves around the application of the Cauchy-Lipschitz theorem. Plugging (4) into (6), we get for any j and k :

$$n'(u | X^k, j) = \frac{n(u | X^k, j) \mu(u | X^k, j)}{\sum_{\ell, g} n(u | X^\ell, g) \mu(u | X^\ell, g)} \quad (22)$$

Introduce the vectors

$$\bar{\mu}(u) = [\mu(u | X^1, f), \dots, \mu(u | X^L, f), \mu(u | X^1, m), \dots, \mu(u | X^L, m)]' \quad (23)$$

$$\bar{n}(u) = [n(u | X^1, f), \dots, n(u | X^L, f), n(u | X^1, m), \dots, n(u | X^L, m)]' \quad (24)$$

A stacked version of (22) is given by:

$$\bar{n}'(u) = g(u, \bar{n}(u)) \quad (25)$$

with:

$$g(u, \bar{n}(u)) = \frac{\bar{n}(u) \cdot * \bar{\mu}(u)}{\langle \bar{n}(u), \bar{\mu}(u) \rangle} \quad (26)$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidian scalar product and for any two vectors V_1 and V_2 of same dimension, $V_1 \cdot * V_2$ is the vector where any element i is the product of the elements i of V_1 and V_2 .

The equation (25) is a first-order differential equation. The denominators of all elements of $g(\cdot, \cdot)$ are strictly positive on $\tilde{\Phi} = (0, 1] \times [0, n(X^1, f)] \times \dots \times [0, n(X^L, f)] \times [0, n(X^1, m)] \times \dots \times [0, n(X^L, m)]$ where $n(X^\ell, j)$ is the measure of gender- j workers with characteristics X^ℓ . This is because there is a constant $c > 0$ such that $\mu(u | X^\ell, m) > c$ and $\mu(u | X^\ell, f) > c$ for all ℓ and all $u \in (0, 1]$. As $\mu(\cdot | X^\ell, m)$ and $\mu(\cdot | X^\ell, f)$ are C^1 on $(0, 1]$ for all ℓ , it is then straightforward to show that $g(\cdot, \cdot)$ is C^1 on $\tilde{\Phi}$. This yields that on any compact set $[\varepsilon, 1] \times [0, n(X^1, f)] \times \dots \times [0, n(X^L, f)] \times [0, n(X^1, m)] \times \dots \times [0, n(X^L, m)]$, $g(\cdot, \cdot)$ is Lipschitzienne and (25) has a unique solution for $\bar{n}(\cdot)$ on $[\varepsilon, 1]$. As this is true for ε arbitrarily close to zero, (25) has a unique solution for $\bar{n}(\cdot)$ on $(0, 1]$.

7.2 Appendix B: consistency

The proof of consistency can be decomposed into three stages:

- We first consider a setting where $\beta(\cdot)$ is constant. This situation corresponds to successive sampling without replacement across ranks, the individuals being drawn with varying probabilities across ranks. Whereas Rosén (1972) establishes consistency when sampling probabilities are constant across successive samples, we extent his results when these probabilities vary.
- We extend the results to the case where $\beta(\cdot)$ is piece-wise constant since proving consistency amounts to doing so for every piece of the function.
- We extend again the results to the case where $\beta(\cdot)$ is continuous function after using a piece-wise constant approximation of it.

We consider in the proof than gender enters the set of explanatory variables X_i to simplify the notations.

1/ Case of constant function $\beta(\cdot)$

We consider a population N , where individuals i have inequal probabilities to be sampled, p_i . In our setting, the individual probability to be sampled depends on her characteristics and is proportional to $X_i\beta$, since β is here constant.

$$p_{X_i}^N = \frac{\exp(X_i\beta)}{\sum_{X'} N_{X'} \exp(X'\beta)}$$

The subscript N refers in this notation to the population at risk, i-e in which individuals can be drawn. Note that

$$\lim_N N p_{X_i}^N = \frac{\exp(X_i \beta)}{\sum_{X'} n_{X'} \exp(X' \beta)} = \frac{\exp(X_i \beta)}{E_{X'} \exp(X' \beta)} \equiv p_{X_i}$$

which we refer as the asymptotic probability for an individual with characteristics X_s to be sampled. N_X is the number of individuals who have characteristics X in the population. Let n be the number of individuals to be sampled, they would occupy the n first jobs in our setting. For instance, if we define $n = \lfloor Nu \rfloor$, these individuals occupy the jobs up to the rank u .

Following Rosén (1972), we define the function $t(\cdot)$ such that

$$N - y = \sum_{i=1}^N \exp(-p_i t(y)) \quad (27)$$

Adapting this definition to our setting, we define the function $t_N(v)$ such that

$$1 - v = \sum_X \frac{N_X}{N} \exp(-N p_X^N t_N(v)) \equiv E_X^N [\exp(-N p_X^N t_N(v))]$$

where $E_X^N(\cdot)$ is the empirical expectation according to X in the population of size N . $t_N(\cdot)$ is increasing over $[0, 1[$, $t_N(0) = 0$. and $\lim_{v \rightarrow 1} t_N(v) = +\infty$. We have $t(Nv) = t_N(v)/N$. We are going to use this relationship to adapt Rosén's result to our setting.

Note that the function $t_N(\cdot)$ can also be defined over $[0, 1[$ by the following differential equation:

$$t'_N(v) = \frac{1}{E_X^N [N p_X^N \exp(-N p_X^N t_N(v))]}$$

with the initial condition $t_N(0) = 0$.

We also define the function $t_\infty(v)$ such that

$$1 - v = E_X \exp(-p_X t_\infty(v))$$

Lemma 1: $t_N(v)$ almost-surely converges to $t_\infty(v)$ for any v in $[0, 1]$. Moreover

$$N^{\frac{1}{2}} (t_N(v) - t_\infty(v)) \rightsquigarrow N \left(0, V_X \left(p_X t_\infty(v) + \frac{\exp(-p_X t_\infty(v))}{1 - v} \right) \right)$$

Proof : We can decompose the expression that gives t_N into

$$\begin{aligned}
& E_X^N [\exp(-Np_X^N t_N(v))] \\
&= E_X [\exp(-Np_X^N t_N(v))] + \sum_X \left(\frac{N_X}{N} - n_X \right) \exp(-Np_X^N t_N(v)) \\
&= E_X [\exp(-p_X t_N(v)) \exp(-(Np_X^N - p_X) t_N(v))] + \sum_X \left(\frac{N_X}{N} - n_X \right) \exp(-Np_X^N t_N(v)) \\
&= E_X [\exp(-p_X t_N(v)) [1 - (Np_X^N - p_X) t_N(v) + O(N^{-1})]] + \sum_X \left(\frac{N_X}{N} - n_X \right) \exp(-Np_X^N t_N(v)) \\
&= E_X [\exp(-p_X t_\infty(v))]
\end{aligned}$$

Hence,

$$\begin{aligned}
& E_X [\exp(-p_X t_N(v))] - E_X [\exp(-p_X t_\infty(v))] \\
&= \sum_X n_X [-(Np_X^N - p_X) \exp(-p_X t_N(v)) t_N(v) + O(N^{-1})] + \sum_X \left(\frac{N_X}{N} - n_X \right) \exp(-Np_X^N t_N(v))
\end{aligned}$$

Hence, since $\lim_{N \rightarrow \infty} Np_X^N = p_X$, $\lim_{N \rightarrow \infty} \frac{N_X}{N} = n_X$, for any $v \in [0, 1[$, $\lim_{N \rightarrow \infty} E_X [\exp(-p_X t_N(v))] = E_X [\exp(-p_X t_\infty(v))] = 1 - v$. Since the function $f(x) = E_X [\exp(-p_X x)]$ is continuous and monotonous, $\lim_{N \rightarrow \infty} t_N(v) = t_\infty(v)$. Moreover, it is possible to derive the asymptotic distribution of $t_N(v)$.

Note that

$$\begin{aligned}
Np_X^N - p_X &= \frac{\exp(X\beta)}{\sum_{X'} \frac{N_{X'}}{N} \exp(X'\beta)} - \frac{\exp(X\beta)}{\sum_{X'} n_{X'} \exp(X'\beta)} \\
&= \frac{\exp(X\beta)}{\sum_{X'} \frac{N_{X'}}{N} \exp(X'\beta) \sum_{X'} n_{X'} \exp(X'\beta)} \sum_{X'} \left(n_{X'} - \frac{N_{X'}}{N} \right) \exp(X'\beta) \\
&= Np_X^N \sum_{X'} \left(n_{X'} - \frac{N_{X'}}{N} \right) p_{X'} \\
&= p_X \sum_{X'} \left(n_{X'} - \frac{N_{X'}}{N} \right) p_{X'} + O(N^{-1})
\end{aligned}$$

Hence,

$$\begin{aligned}
& E_X [\exp(-p_X t_N(v))] - E_X [\exp(-p_X t_\infty(v))] \\
&= E_X [\exp(-p_X t_N(v))] - E_X^N [\exp(-N p_X^N t_N(v))] \\
&= E_X [\exp(-p_X t_N(v)) (1 - \exp(-(N p_X^N - p_X) t_N(v)))] + \sum_X \left(\frac{N_X}{N} - n_X \right) \exp(-N p_X^N t_N(v)) \\
&= \sum_{X'} n_{X'} [-(N p_X^N - p_X) \exp(-p_X t_N(v)) t_N(v) + O(N^{-1})] + \sum_X \left(\frac{N_X}{N} - n_X \right) \exp(-N p_X^N t_N(v)) \\
&= \sum_X \left(\frac{N_X}{N} - n_X \right) [p_X t_\infty(v) E_{X'}(p_{X'} \exp(-p_{X'} t_\infty(v))) + \exp(-p_X t_\infty(v))] + o(N^{-1/2})
\end{aligned}$$

Hence

$$N^{1/2} (E_X [\exp(-p_X t_N(v))] - E_X [\exp(-p_X t_\infty(v))]) \rightarrow N(0, V_X(p_X t_\infty(v)(1-v) + \exp(-p_X t_\infty(v))))$$

Hence,

$$N^{1/2} (t_N(v) - t_\infty(v)) \rightarrow N\left(0, V_X\left(p_X t_\infty(v) + \frac{\exp(-p_X t_\infty(v))}{1-v}\right)\right)$$

applying the delta method to the invert function of $f(x) = E_X[\exp(-p_X x)]$, since $f^{-1'}(y) = 1/f'(f^{-1}(y)) = 1/E_X[\exp(-p_X f^{-1}(y))]$.

This result comes from the following:

$$\begin{aligned}
& V_{as} \left(\sum_{X'} \left(n_{X'} - \frac{N_{X'}}{N} \right) f(X') \right) \\
&= N^{-1} [\dots f(X'_k) \dots] \begin{bmatrix} n_{X_1} (1 - n_{X_1}) & & & \\ -n_{X_k} n_{X_{k'}} & n_{X_k} (1 - n_{X_k}) & & \\ -n_{X_1} n_{X_K} & & n_{X_K} (1 - n_{X_K}) & \\ & & & \dots \end{bmatrix} \begin{bmatrix} \dots \\ f(X'_k) \\ \dots \end{bmatrix} \\
&= N^{-1} [\dots f(X'_k) \dots] \begin{bmatrix} \dots & & & \\ n_k f(X'_k) - \sum_{k'} n_{X_k} n_{X_{k'}} f(X'_{k'}) & & & \\ \dots & & & \dots \end{bmatrix} \\
&= N^{-1} \left(\sum_{X'} n_{X'} f(X')^2 - \left(\sum_{X'} n_{X'} f(X') \right)^2 \right) \\
&= N^{-1} V_{X'} [f(X')]
\end{aligned}$$

Let's come back to Rosen's setting. A quantity a_i is associated to each individual. Let $Z_n = \sum_{i \in n} a_i$ be the sum of these quantities over the sampled individuals. From Rosen, we get that

$$EZ_n = \sum_{i=1}^N a_i \left(1 - \exp \left(-p_i^N t_N \left(\frac{n}{N} \right) \right) \right) + o \left(N^{\frac{1}{2}} \right)$$

Consider the specific case where a_i is a dummy indicating if i has the characteristics X , then Z_n is the number of sampled individuals with characteristics X in the first n jobs, i-e $M_{n,N}^X$.

Hence, from Rosén's theorem 3.1, we get that

$$EM_{n,N}^X = N_X \left(1 - \exp \left(-p_X^N t_N \left(\frac{n}{N} \right) \right) \right) + r_E(n) N^{\frac{1}{2}}$$

where $\lim_{N \rightarrow \infty} \max_{\tau_1 N \leq n \leq \tau_2 N} |r_E(n)| = 0$ for every $0 < \tau_1 < \tau_2 < 1$. Hence,

$$\begin{aligned} \frac{EM_{N-\lfloor Nu \rfloor, N}^X}{N} &= \frac{N_X}{N} \left(1 - \exp \left(-p_X^N t_N \left(1 - \frac{\lfloor Nu \rfloor}{N} \right) \right) \right) + r_E(n) N^{-\frac{1}{2}} \\ &= n_X (1 - \exp(-p_X t_\infty (1 - u))) + r_E(n) N^{-\frac{1}{2}} \end{aligned}$$

From the lemma 1 above, we have that $\frac{N_X}{N} (1 - \exp(-p_X^N t_N (1 - u))) - n_X (1 - \exp(-p_X t_\infty (1 - u)))$ converges to 0 like $N^{-1/2}$. We also have that $\left| u - \frac{\lfloor Nu \rfloor}{N} \right| \leq \frac{1}{N}$. Hence $\frac{N_X}{N} (1 - \exp(-p_X^N t_N (1 - \frac{\lfloor Nu \rfloor}{N}))) - \frac{N_X}{N} (1 - \exp(-p_X^N t_N (1 - u)))$ converges to 0 like N^{-1} . Finally, in this setting, $n = N - \lfloor Nu \rfloor$. Hence, we can use the result of Rosen with $\tau_1 = u$ and $\tau_2 = 1 - \varepsilon$.

In the end, we have that

$$\frac{EM_{N-\lfloor Nu \rfloor, N}^X}{N} = n_X (1 - \exp(-p_X t_\infty (1 - u))) + o \left(N^{-1/2} \right)$$

Define now $\bar{n}(u; X) = n_X \exp(-p_X t_\infty (1 - u))$. We have that $\bar{n}(1; X) = n_X$.

$$\begin{aligned} \bar{n}'(u; X) &= n_X p_X t_\infty' (1 - u) \exp(-p_X t_\infty (1 - u)) \\ &= \frac{p_X n_X \exp(-p_X t_\infty (1 - u))}{\sum_{X'} p_{X'} n_{X'} \exp(-p_{X'} t_\infty (1 - u))} \\ &= \frac{p_X \bar{n}(u; X)}{\sum_{X'} p_{X'} \bar{n}(u; X')} \\ &= \frac{\exp(X\beta) \bar{n}(u; X)}{\sum_{X'} \exp(X'\beta) \bar{n}(u; X')} \end{aligned}$$

which corresponds to the definition of the equation that defines $n(u; X)$.

since

$$\begin{aligned}
V \left(M_{N-\lfloor Nu \rfloor, N}^X \right) &= \sum_{s=1}^N \left(1_{s \in X} - \frac{\sum_{r=1}^N p_r 1_{r \in X} e^{-p_r t_N(\lfloor Nu \rfloor / N)}}{\sum_{r=1}^N p_r e^{-p_r t_N(\lfloor Nu \rfloor / N)}} \right)^2 \left(1 - e^{-p_s t_N(\lfloor Nu \rfloor / N)} \right) e^{-p_s t_N(\lfloor Nu \rfloor / N)} + r_\sigma(n) \\
&= \sum_{s=1}^N \left[1_{s \in X} - 2 \cdot 1_{s \in X} \frac{N_X p_X^N e^{-p_X^N t_N(\lfloor Nu \rfloor / N)}}{\sum_{X', N_{X'} p_{X'}^N e^{-p_{X'}^N t_N(\lfloor Nu \rfloor / N)}} \right. \\
&\quad \left. + \left(\frac{N_X p_X^N e^{-p_X^N t_N(\lfloor Nu \rfloor / N)}}{\sum_{X', N_{X'} p_{X'}^N e^{-p_{X'}^N t_N(\lfloor Nu \rfloor / N)}} \right)^2 \right] \left(1 - e^{-p_s t_N(\lfloor Nu \rfloor / N)} \right) e^{-p_s t_N(\lfloor Nu \rfloor / N)} + r_\sigma(n) \\
&= N_X \left[1 - 2 \frac{p_X N_X e^{-p_X t_N(\lfloor Nu \rfloor / N)}}{\sum_{X', p_{X'} N_{X'} e^{-p_{X'} t_N(\lfloor Nu \rfloor / N)}} \right] \left(1 - e^{-p_X t_N(\lfloor Nu \rfloor / N)} \right) e^{-p_X t_N(\lfloor Nu \rfloor / N)} \\
&\quad + \left(\frac{N_X p_X e^{-p_X t_N(\lfloor Nu \rfloor / N)}}{\sum_{X', N_{X'} p_{X'} e^{-p_{X'} t_N(\lfloor Nu \rfloor / N)}} \right)^2 \sum_{X'} N_{X'} \left(1 - e^{-p_{X'} t_N(\lfloor Nu \rfloor / N)} \right) e^{-p_{X'} t_N(\lfloor Nu \rfloor / N)} + r_\sigma(\lfloor Nu \rfloor)
\end{aligned}$$

Hence

$$\begin{aligned}
V \left(\frac{M_{N-\lfloor Nu \rfloor, N}^X}{N} \right) &= \frac{1}{N} \left\{ \begin{aligned} &n_X \left[1 - 2 \frac{p_X n(u; X)}{\sum_{X', p_{X'} n(u; X')} \right] \frac{n(u; X)}{n_X} \left(1 - \frac{n(u; X)}{n_X} \right) \\ &+ \left(\frac{p_X n(u; X)}{\sum_{X', p_{X'} n(u; X')} \right)^2 \sum_{X'} n(u; X') \left(1 - \frac{n(u; X')}{n_{X'}} \right) \end{aligned} \right\} \\
&\quad + \frac{1}{N} r_\sigma(\lfloor Nu \rfloor) + o\left(\frac{1}{N}\right)
\end{aligned}$$

where $r_\sigma(n)$ tends to 0 as n becomes large. As a result, $\frac{M_{N-\lfloor Nu \rfloor, N}^X}{N} = \frac{EM_{N-\lfloor Nu \rfloor, N}^X}{N} + O(N^{-1/2})$, and we deduce that

$$\frac{N_{\lfloor Nu \rfloor}^X}{N} = \frac{N_X - M_{N-\lfloor Nu \rfloor, N}^X}{N} = n(u; X) + O(N^{-1/2})$$

Note that when u comes close to zero, there might be an issue of the number of observations that become too low for the asymptotics considerations to apply. This is the reason why we should not consider ranks too low with respect to the number of available observations.

2/ Case of a piece-wise constant function $\beta(\cdot)$

The knots are denoted $\varepsilon < u_1 < \dots < u_Z < 1$. The exact same logic can be applied between over the interval $[u_Z, 1]$. We have that

$$\lim_N E \frac{N_{\lfloor Nu \rfloor}^X}{N} = n(u; X), \forall u \in]u_Z, 1]$$

where

$$n'(u; X) = \frac{\exp(X\beta(1)) n(u; X)}{\sum_{X'} \exp(X'\beta(1)) n(u; X')}$$

with $n(1; X) = n_X$. This proves the existence and unicity of the function $n(u; X)$ on the interval $]u_Z, 1]$.

Assume now that this relationship holds up to the rank u_z . Then, from the setting above, we have that

$$EM_{n, \lfloor Nu_z \rfloor}^X = N_{\lfloor Nu_z \rfloor}^X \left(1 - \exp \left(-p_{X \lfloor Nu_z \rfloor}^{z} \left(\frac{n}{\lfloor Nu_z \rfloor} \right) \right) \right) + o \left(N^{\frac{1}{2}} \right)$$

since individuals drawn in the ranks between u_z and u_{z-1} are drawn with the same individual probabilities from the population still available at rank $\lfloor Nu_z \rfloor$:

$$p_{X_s}^{\lfloor Nu_z \rfloor} = \frac{\exp(X_s \beta(u_z))}{\sum_{X'} N_{X'}^{\lfloor Nu_z \rfloor} \exp(X' \beta(u_z))}$$

and for any rank between u_{z-1} and u_z , we have that

$$\frac{N_{\lfloor Nu \rfloor}^X}{N} = n(u; X) + O \left(N^{-1/2} \right), \forall u \in]u_{z-1}, u_z]$$

where $n(u; X)$ is defined from the following system of equations:

$$n(u_z; X) = \lim_{u > u_z} n(u; X)$$

as the initial condition, and

$$n'(u; X) = \frac{\exp(X \beta(u_z)) n(u; X)}{\sum_{X'} \exp(X' \beta(u_z)) n(u; X')}, \forall u \in]u_{z-1}, u_z]$$

By recurrence, the function $n(u; X)$ is defined over the interval $[\varepsilon, 1]$ and corresponds to the expectation of the share of individuals with characteristics X who are still available for the job of rank u .

3/ Case of a continuous function $\beta(\cdot)$

Consider a step-wise approximation β^Z of β over the ranks, such that $\beta^Z(u) = \sum_{z=1}^Z \beta\left(\frac{z}{Z}\right) 1\left\{\frac{z-1}{Z} \leq u < \frac{z}{Z}\right\}$. We have that $\max_u \left| \beta^Z(u) - \beta(u) \right| \leq \frac{1}{Z} \max_u |\beta'(u)|$, from the mean value theorem.

From the previous result, we can define $n_Z(u; X)$ as the measure of individuals with characteristics X still available for the u -rank job, with an assignment mechanism defined by β^Z and $N_{\lfloor Nu \rfloor}^{X,z}$ the random variable that corresponds to the number of individuals with characteristics X still available for rank u .

We prove by recurrence that for any X, z ,

$$Diff_z^{X,Z} = \frac{N_{\lfloor N \frac{z}{Z} \rfloor}^X - N_{\lfloor N \frac{z}{Z} \rfloor}^{X,Z}}{N} = O \left((N/Z)^{-1/2} \right) + O \left(Z^{-1/2} \right) \quad (28)$$

[DETAILS TO BE ADDED]

7.3 Appendix C: theoretical fundations of the simulation approach

The finite discrete counterpart of the differential equation verified by the measures of available workers (6) can be rewritten in vector form piling up terms in the (X, j) dimension as:

$$\vec{N}^* (u^k) = \vec{N}^* (u^{k+1}) - \vec{D}_{k+1} \quad (29)$$

where $\vec{N}^* (u) = [N^* (u | X^1, j_1), \dots, N^* (u | X^L, j_2)]'$ and $\vec{D}_k = [D_k (X^1, j_1), \dots, D_k (X^L, j_2)]'$ with $\{j_1, j_2\} = \{f, m\}$. It is straightforward to show recursively that:

$$E [\vec{N}^* (u^k)] = \vec{N} - E (\vec{M}_{k+1}) \quad (30)$$

where $\vec{M}_k = \sum_{\ell=k}^N \vec{D}_\ell$ and $\vec{N} = [N (X^1, j_1), \dots, N (X^L, j_2)]'$ where $N (X, j)$ is the number of gender- j workers with characteristics X in the sample. We need a strategy to estimate the second right-hand side term. We have:

$$E (\vec{M}_k) = E_{\vec{M}_k, \vec{M}_{k+1}, \dots, \vec{M}_N} (\vec{M}_k) \quad (31)$$

The expectation $E (\vec{M}_k)$ can thus be computed by simulation, averaging across iterations the values of \vec{M}_k obtained when drawing values of $\vec{M}_k, \vec{M}_{k+1}, \dots,$ and \vec{M}_N which joint law verifies:

$$P (\vec{M}_k = \vec{m}_k, \vec{M}_{k+1} = \vec{m}_{k+1}, \dots, \vec{M}_N = \vec{m}_N) \quad (32)$$

$$= \left[\prod_{\ell=k+1}^{N-1} P (\vec{M}_k = \vec{m}_k | \vec{M}_{k+1} = \vec{m}_{k+1}) \right] P (\vec{M}_N = \vec{m}_N) \quad (33)$$

Draws can thus be made first drawing in the law of \vec{M}_N , and then sequentially in the law of \vec{M}_k conditionally on the simulated value of \vec{M}_{k+1} denoted \vec{m}_{k+1}^s (Gourieroux and Monfort, 1995). The law of $\vec{M}_N = \vec{D}_N$ is simply that of a multinomial logit which probabilities are given by:

$$P (1 | X, j) = \frac{N (X, j) \exp [X \beta_j^* (1)]}{\sum_{\ell, g} N (X^\ell, g) \exp [X^\ell \beta_g^* (1)]} \quad (34)$$

where $N (X, j)$ is the number of gender- j workers with characteristics X in the sample. Denote by $\Omega_j^* (u, X)$ the random set that contains all available gender- j workers with characteristics X at rank u and $\vec{\Omega}^* (u) = \{\Omega_{j_1}^* (u, X^1), \dots, \Omega_{j_2}^* (u, X^L)\}$. The law of $\vec{M}_k | \vec{M}_{k+1} = \vec{m}_{k+1}^s$ is simple because \vec{m}_{k+1}^s contains all the information necessary to determine the realization of $\vec{\Omega}^* (u)$ that we denote $\vec{\Omega}^s (u)$. We have:

$$P (\vec{M}_k = \vec{m}_k | \vec{M}_{k+1} = \vec{m}_{k+1}^s) = P [\vec{D}_k = \vec{m}_k - \vec{m}_{k+1}^s | \vec{\Omega}^* (u^k) = \vec{\Omega}^s (u^k)] \quad (35)$$

where $\vec{m}_k - \vec{m}_{k+1}^s$ is a vector where there is only one element that takes the value one and this occurs at the position associated to the characteristics (X, j) of the worker that gets the job, and other elements of the vector take the value zero. The probability that the element corresponding to a given (X, j) takes the value one (and other elements take value zero) is given by:

$$P^s(u^k | X, j) = \frac{N^s(u^k | X, j) \exp[X\beta_j^*(u^k)]}{\sum_{\ell, g} N^s(u^k | X^\ell, g) \exp[X\beta_g^*(u^k)]} \quad (36)$$

where $N^s(u | X, j) = \text{Card } \Omega_j^s(u, X)$ with $\Omega_j^s(u, X)$ the set of available gender- j workers with characteristics X .

For a given simulation iteration, we first draw for rank 1 in the law of a multinomial logit using formula (34), and we then draw sequentially for ranks $u^k = (k - 1) / (N - 1)$ with $k = N - 1, \dots, 1$ in the laws of multinomial logits using formula (36). A simulated value of \vec{M}_k is denoted by \vec{M}_k^s . A consistent estimator of $E[\vec{N}^*(u^k)]$ when the number of simulations tends to infinity is then given by $\vec{N} - \sum_{s=1}^S \vec{M}_{k+1}^s / S$.

References

- [1] Aigner D. and G. Cain (1977), “Statistical Theories of Discrimination in Labor Markets”, *Industrial and Labor Relations Review*, 30(2), pp. 175-187.
- [2] Albrecht J., Björklund A. and S. Vroman (2003), “Is There a Glass Ceiling in Sweden?”, *Journal of Labour Economics*, 21(1), pp. 145-177.
- [3] Andersen P. and R. Gill (1982), “Cox’s regression model for counting processes: a large sample study”, *The annals of statistics*, 10(4), pp. 1100-1120.
- [4] Bradley J., Postel-Vinay F. and H. Turon (2013), “Public Sector Wage Policy and Labor Market Equilibrium: A Structural Model”, Working Paper.
- [5] Chernozhukov V., Fernández-Val I. and B. Melly (2013), “Inference on counterfactual distributions”, *Econometrica*, 81(6), pp. 2205-2268.
- [6] Dagsvik J. (1994), “Discrete and Continuous Choice, Max-Stable Processes, and Independence from Irrelevant Attributes”, *Econometrica*, 62(5), pp. 1179-1205.
- [7] Daussin-Benichou J., Koubi M., Leduc A. and B. Marc (2014), “Les carrières salariales dans le public et le privé”, in *Emploi et salaires*, édition 2014, Insee, pp 47-60.
- [8] Doeringer, P. and M. Piore (1985). *Internal labor markets and manpower analysis*, ME Sharpe, 248p.
- [9] Dorothee O. and E. Baradji E. (2014), “L’emploi dans les trois versants de la fonction publique”, *Point stat*, 15, DGAFP, Ministère de la décentralisation et de la fonction publique.
- [10] Dorothee O., Le Faller P. and M. Treppoz (2013), “Les principales caractéristiques des agents de la fonction publique au 31 décembre 2011”, *Point stat*, 11, DGAFP, Ministère de la décentralisation et de la fonction publique.
- [11] Firpo, N. Fortin and T. Lemieux (2009), “Unconditional quantile regressions”, *Econometrica*, 77(3), pp. 953-973.
- [12] Fortin N. (2008), “The Gender Wage Gap among Young Adults in the United States The importance of Money versus People”, *Journal of Human Resources*, 43(4), pp.884-918 .
- [13] Fortin N., Lemieux T. and N. Firpo (2011) , “Decomposition methods in economics”, in *Handbook of Labor Economics*, Ashenfelter O. and D. Card, Volume 4, Part A, pp. 1–102.

- [14] Gayle G.E. and L. Golan (2012), “Estimating a Dynamic Adverse-Selection Model: Labour-Force Experience and the Changing Gender Earnings Gap 1968–1997”, *The Review of Economic Studies*, 79 (1), pp. 227-267.
- [15] Gobillon L., Meurs D. and S. Roux (2015), “Estimating gender differences in access to jobs”, *Journal of Labour Economics*, 33(2), pp. 317-363.
- [16] Gregory R. and J. Borland (1999), “Recent developments in public sector labor markets”, in *Handbook of Labor Economics*, Ashenfelter O. and D. Card, Volume 3, Part C, pp. 3573–3630.
- [17] Machado J. and J. Mata (2005), “Counterfactual decomposition of changes in wage distributions using quantile regression”, *Journal of Applied Econometrics*, 20, pp. 445–465.
- [18] Postel-Vinay F. and H. Turon (2007), “The Public Pay Gap in Britain: Small Differences That (Don’t?) Matter”, *The Economic Journal*, 117(523), pp. 1460-1503.
- [19] Lucifora C. and D. Meurs (2006), “The Public Sector Pay Gap in France, Great Britain and Italy”, *Review of Income and Wealth*, 52(1), pp. 43–59.
- [20] Melly (2005a), “Decomposition of differences in distribution using quantile regression”, *Labour Economics*, 12, pp. 577–590.
- [21] Melly (2005b), “Public-private sector wage differentials in Germany: Evidence from quantile regression”, *Empirical Economics*, 30, pp. 505–520.
- [22] Meurs D. and S. Ponthieux (2015), “Gender inequality”, in *Handbook of Income Distribution*, Atkinson A. and F. Bourguignon, vol. 2, Elsevier, Chapter 12, pp. 981-1146.
- [23] Rosén (1972), “Asymptotic Theory fo Successive Sampling with Varying Probabilities witout Replacement, I”, *The Annals of Mathematical Statistics*, 43(2), pp. 373-397.
- [24] Rothe C. (2012), “Partial Distributional Policy Effects”, *Econometrica*, 80, pp. 2269–2301.
- [25] Sattinger M. (1975), “Comparative Advantage and the Distribution of Earnings and Abilities”, *Econometrica*, 43, 455-468.
- [26] Sattinger M. (1993), “Assignment Models of the Distribution of Earnings”, *Journal of Economic Literature*, 31(2), pp. 831-880.
- [27] Teulings C. (1995), “The Wage Distribution in a Model of the Assignment of Skills to Jobs”, *The Journal of Political Economy*, 103(2), pp. 280-315.

Table 1: Descriptive statistics on wages by gender

	Mean	Std Dev	P5	P10	P25	Median	P75	P90	P95	N
<u>All sectors</u>										
All	77.7	65.0	40.1	43.4	50.5	62.8	84.7	122.1	158.9	55,881
Males	83.7	73.3	41.8	45.3	53.2	66.3	90.7	133.4	175.6	34,745
Females	67.9	46.6	38.5	41.2	47.3	56.9	75.6	102.7	129.4	21,136
% diff. F-M	-18.9%	-36.5%	-7.7%	-9.1%	-11.1%	-14.1%	-16.7%	-23.0%	-26.4%	
<u>Public sector</u>										
All	70.6	32.4	41.8	45.3	51.1	62.2	79.6	102.2	126.6	9,732
Males	76.2	36.0	44.4	47.9	54.8	67.2	84.7	112.6	139.7	4,781
Females	65.2	27.4	40.1	43.5	48.9	57.2	73.3	93.7	115.3	4,951
% diff. F-M	-14.4%	-23.8%	-9.8%	-9.2%	-10.9%	-14.9%	-13.4%	-16.7%	-17.4%	
<u>Private sector</u>										
All	79.2	69.9	39.8	43.1	50.2	63.0	86.4	126.7	165.8	46,149
Males	84.9	77.6	41.4	45.0	52.9	66.1	92.1	137.1	180.3	29,964
Females	68.7	51.0	38.2	40.7	46.7	56.8	76.3	105.9	133.3	16,185
% diff. F-M	-19.1%	-34.3%	-7.6%	-9.4%	-11.7%	-14.0%	-17.2%	-22.7%	-26.1%	

Note: Statistics are computed on the daily wage. The % difference is the female value minus the male value divided by the male value. Std Dev: Standard Deviation, PX: Xth percentile.

Table 2: Descriptive statistics on explanatory variables by gender

	All sectors			Public sector			Private sector			
	All	Males	Females	All	Males	Females	All	Males	Females	% diff.
Female	0.378			0.509			0.351			
Diploma										
<High-School	0.706	0.731	0.665	0.728	0.744	0.712	0.701	0.729	0.651	-10.7%
≥High-School	0.294	0.269	0.335	0.272	0.256	0.288	0.299	0.271	0.349	+28.7%
Children										
0	0.330	0.330	0.329	0.334	0.332	0.337	0.329	0.330	0.327	-0.9%
1 or 2	0.548	0.535	0.570	0.542	0.538	0.545	0.550	0.534	0.578	+8.2%
≥3	0.122	0.135	0.100	0.124	0.130	0.118	0.122	0.136	0.095	-30.1%
Age										
31-40	0.316	0.327	0.298	0.245	0.274	0.218	0.331	0.336	0.323	-3.9%
41-50	0.366	0.367	0.364	0.353	0.355	0.351	0.369	0.369	0.369	-0.1%
≥51	0.318	0.306	0.338	0.402	0.371	0.432	0.300	0.296	0.309	+4.5%
Paris region										
Inside	0.235	0.223	0.255	0.234	0.211	0.256	0.235	0.224	0.255	+13.8%
Outside	0.765	0.777	0.745	0.766	0.786	0.744	0.765	0.776	0.745	-4.0%
Job seniority										
≤10 years	0.359	0.353	0.370	0.421	0.413	0.428	0.687	0.684	0.692	+1.2%
>10 years	0.640	0.647	0.630	0.579	0.587	0.572	0.313	0.316	0.308	-2.5%
Part-time experience										
≤7%	0.493	0.585	0.342	0.459	0.600	0.322	0.500	0.582	0.348	-40.2%
>7% and ≤18%	0.232	0.239	0.221	0.202	0.215	0.189	0.239	0.243	0.231	-5.1%
>18%	0.275	0.176	0.437	0.339	0.184	0.489	0.261	0.175	0.421	+141.2%
Work interruption										
≤1 year	0.208	0.211	0.203	0.233	0.237	0.229	0.203	0.207	0.196	-5.6%
>1 and ≤3 years	0.258	0.277	0.227	0.184	0.207	0.162	0.274	0.289	0.246	-14.7%
>3 and ≤6 years	0.265	0.274	0.251	0.239	0.256	0.223	0.271	0.277	0.259	-6.5%
>6 years	0.268	0.237	0.319	0.344	0.300	0.386	0.252	0.227	0.299	+31.7%

Note: Figures in the table correspond to proportions except % difference which is the female value minus the male value divided by the male value.

Table 3: Observed and counterfactual log-wage gaps obtained in different scenarios

	Public sector		Private sector		Observed		Private sector		Public	
	Observed	Public values	Equal values	Private values	Observed	Private values	Equal values	Public values		
Mean										
Males	4.2596	4.2587	4.1985	4.2628	4.2970	4.2964	4.2409	4.2929		
Females	4.1165	4.1173	4.1755	4.1134	4.1295	4.1306	4.2339	4.1371		
F-M Gap	-0.1431	-0.1415	-0.0230	-0.1494	-0.1676	-0.1658	-0.0070	-0.1558		
First decile										
Males	3.8698	3.8715	3.8347	3.8732	3.8091	3.8095	3.7737	3.8026		
Females	3.7729	3.7731	3.7923	3.7715	3.7045	3.7047	3.7462	3.7116		
F-M Gap	-0.0969	-0.0984	-0.0425	-0.1016	-0.1046	-0.1048	-0.0275	-0.0910		
First quartile										
Males	4.0044	4.0038	3.9495	4.0005	3.9675	3.9679	3.9235	3.9707		
Females	3.8893	3.8886	3.9176	3.8879	3.8393	3.8398	3.9013	3.8464		
F-M Gap	-0.1151	-0.1152	-0.0320	-0.1127	-0.1281	-0.1282	-0.0222	-0.1244		
Median										
Males	4.2070	4.2056	4.1430	4.1975	4.1875	4.1893	4.1436	4.1997		
Females	4.0461	4.0448	4.1114	4.0500	4.0393	4.0411	4.1356	4.0286		
F-M Gap	-0.1609	-0.1608	-0.0316	-0.1476	-0.1482	-0.1482	-0.0080	-0.1711		
Last quartile										
Males	4.4388	4.4410	4.3824	4.4509	4.5224	4.5231	4.4564	4.5128		
Females	4.2952	4.2955	4.3719	4.3015	4.3362	4.3405	4.4700	4.3297		
F-M Gap	-0.1435	-0.1454	-0.0105	-0.1494	-0.1862	-0.1825	0.0136	-0.1831		
Last decile										
Males	4.7236	4.7238	4.6255	4.7542	4.9269	4.9258	4.8394	4.8991		
Females	4.5403	4.5451	4.6285	4.5196	4.6607	4.6644	4.8541	4.7201		
F-M Gap	-0.1833	-0.1787	0.0030	-0.2346	-0.2662	-0.2614	0.0147	-0.1790		
Standard Deviation										
Males	0.3584	0.3568	0.3439	0.3668	0.4740	0.4714	0.4530	0.4648		
Females	0.3284	0.3303	0.3569	0.3177	0.4010	0.4070	0.4640	0.4231		
F-M Gap	-0.0300	-0.0265	0.0131	-0.0491	-0.0730	-0.0645	0.0110	-0.0417		

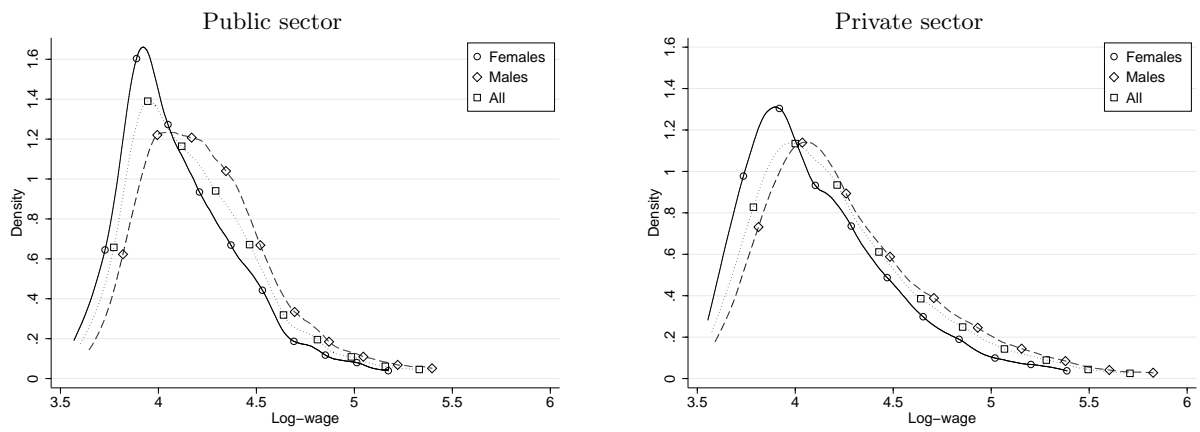
Note: Statistics are computed on the logarithm of daily wage. Column headings mention either that statistics are computed directly from the data (label "Observed") or that they are derived from counterfactuals using the conditional worker values of the public sector, the private sector, or the same conditional worker values for the two gender in the sector that is considered (respective labels "Private values", "Public values", "Equal values").

Table 4: Observed and counterfactual log-wage gaps obtained in different scenarios, hourly wages, part-time workers included

	Observed	Public sector			Private sector			
		Public values	Equal values	Private values	Observed	Private values	Equal values	Public values
Mean								
Males	2.6197	2.6190	2.5642	2.6282	2.6345	2.6301	2.5689	2.6247
Females	2.4811	2.4817	2.5214	2.4750	2.4536	2.4603	2.5435	2.4677
F-M Gap	-0.1385	-0.1373	-0.0427	-0.1532	-0.1809	-0.1698	-0.0255	-0.1571
First decile								
Males	2.2263	2.2272	2.2052	2.2439	2.1529	2.1549	2.1183	2.1404
Females	2.1497	2.1508	2.1540	2.1440	2.0513	2.0534	2.0753	2.0654
F-M Gap	-0.0767	-0.0764	-0.0512	-0.0999	-0.1016	-0.1014	-0.0430	-0.0750
First quartile								
Males	2.3621	2.3615	2.3169	2.3635	2.3042	2.3042	2.2562	2.3038
Females	2.2571	2.2575	2.2741	2.2532	2.1700	2.1698	2.2155	2.1799
F-M Gap	-0.1050	-0.1039	-0.0428	-0.1103	-0.1342	-0.1344	-0.0407	-0.1239
Median								
Males	2.5619	2.5617	2.5017	2.5633	2.5265	2.5291	2.4712	2.5342
Females	2.4065	2.4054	2.4494	2.4049	2.3595	2.3636	2.4374	2.3594
F-M Gap	-0.1553	-0.1563	-0.0523	-0.1585	-0.1670	-0.1655	-0.0338	-0.1748
Last quartile								
Males	2.8053	2.8062	2.7455	2.8195	2.8804	2.8765	2.7875	2.8678
Females	2.6530	2.6541	2.7116	2.6494	2.6563	2.6594	2.7797	2.6573
F-M Gap	-0.1523	-0.1521	-0.0339	-0.1701	-0.2241	-0.2171	-0.0077	-0.2104
Last decile								
Males	3.0905	3.0928	2.9951	3.1151	3.2511	3.2391	3.1519	3.2216
Females	2.9012	2.9009	2.9644	2.8799	2.9916	3.0129	3.1524	3.0440
F-M Gap	-0.1893	-0.1919	-0.0307	-0.2352	-0.2595	-0.2262	0.0006	-0.1776
Standard deviation								
Males	0.3603	0.3571	0.3382	0.3593	0.4560	0.4481	0.4282	0.4463
Females	0.3132	0.3160	0.3414	0.3114	0.3769	0.3927	0.4406	0.3984
F-M Gap	-0.0471	-0.0411	0.0032	-0.0479	-0.0791	-0.0554	0.0124	-0.0480

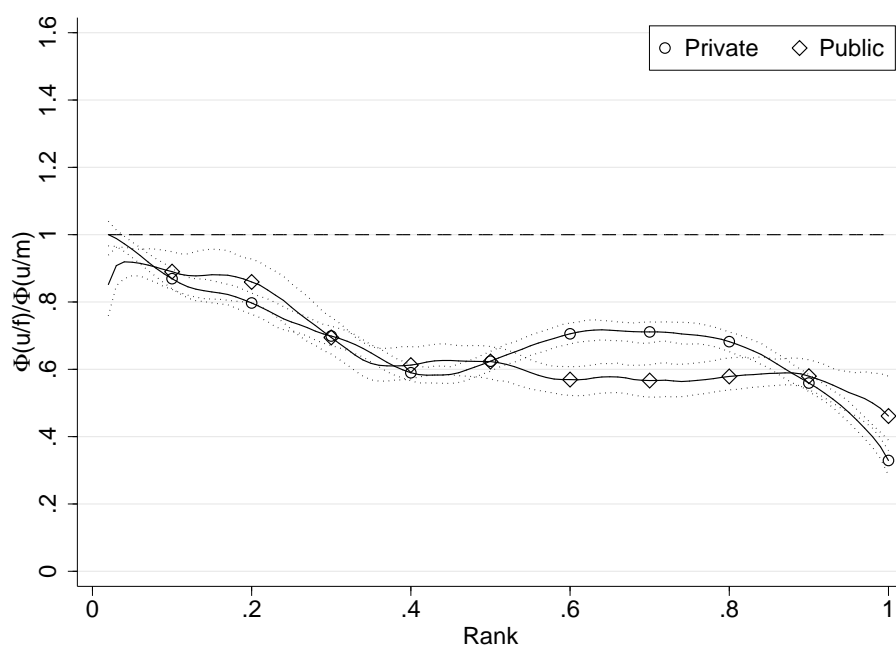
Note: Statistics are computed on the logarithm of hourly wage. Column headings mention either that statistics are computed directly from the data (label "Observed") or that they are derived from counterfactuals using the conditional worker values of the public sector, the private sector, or the same conditional worker values for the two gender in the sector that is considered (respective labels "Private values", "Public values", "Equal values").

Figure 1: Gender log-wage distributions in the public and private sectors



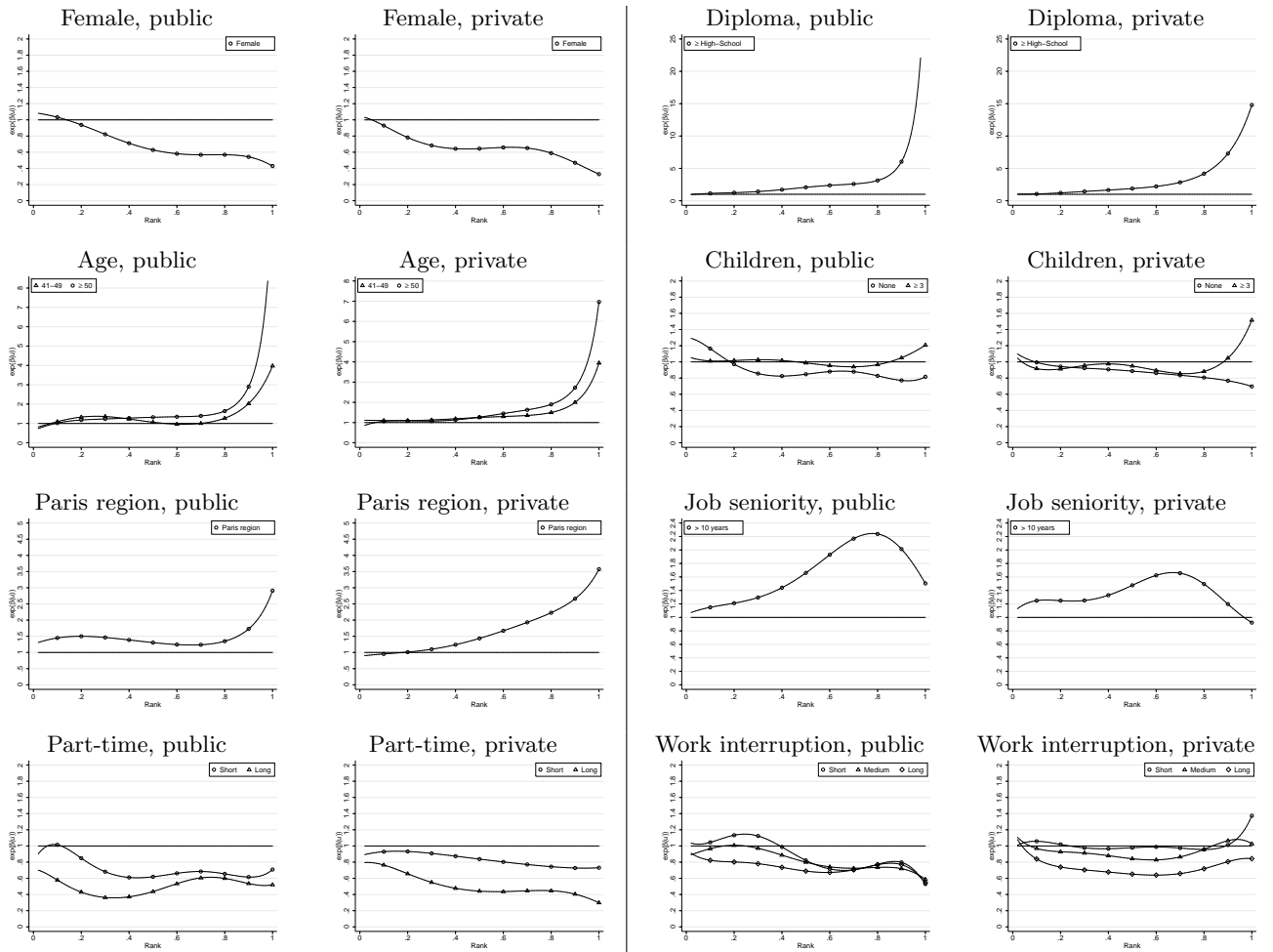
Note: Densities are computed using the logarithm of daily wage.

Figure 2: Gender probability ratio of getting a given job in the public and private sectors



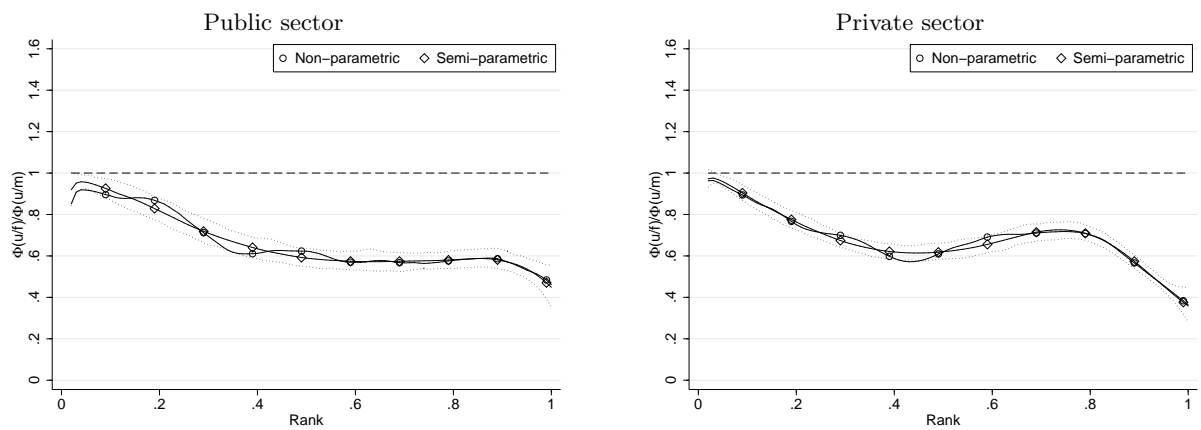
Note: Confidence intervals at the 5% level obtained by bootstrap using 100 replications are reported in dotted lines.

Figure 3: Exponentiated effects of dummy variables on conditional worker values



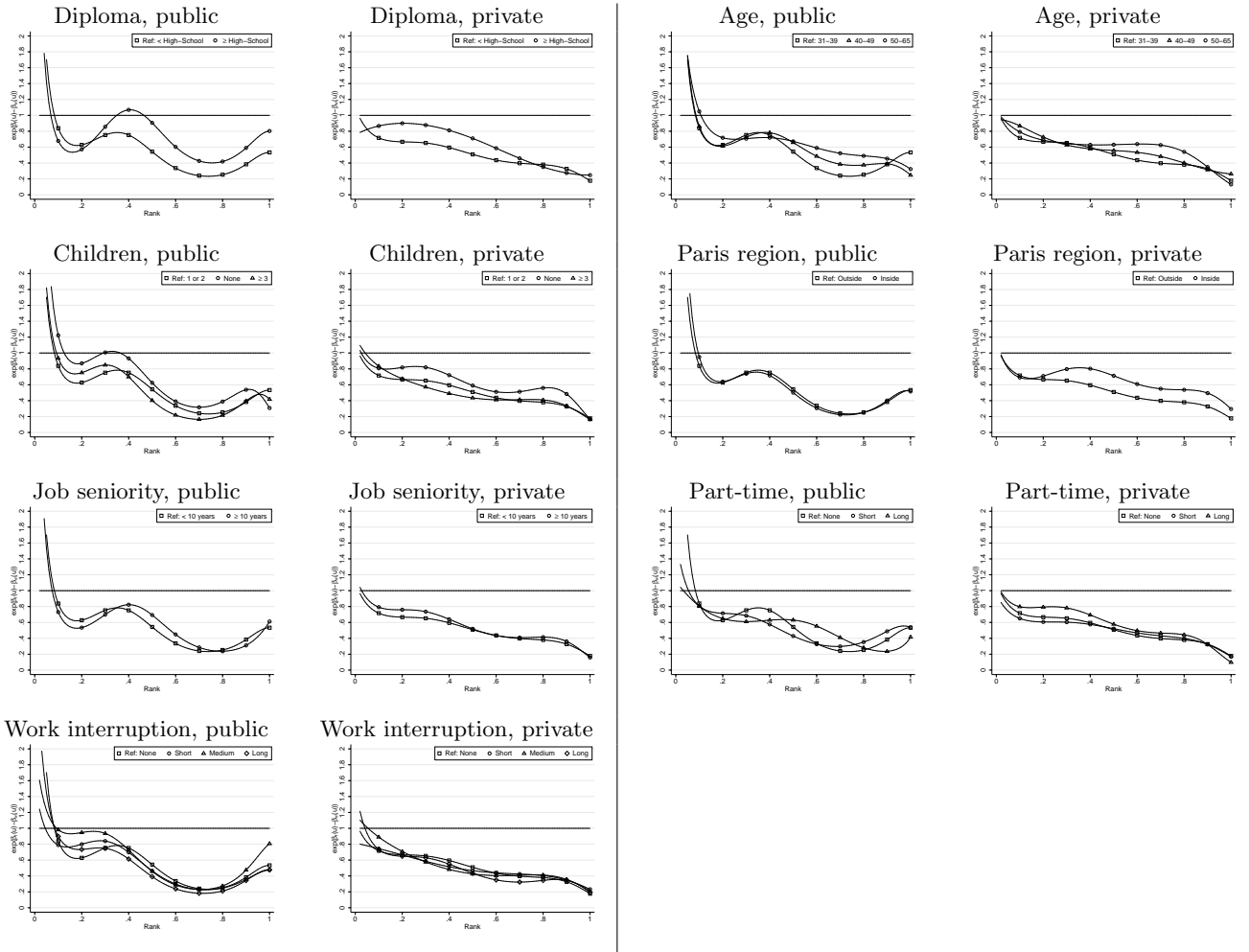
Note: The graph title gives the characteristic for which the exponentiated coefficients are graphed, as well as the sector.

Figure 4: Non-parametric and semi-parametric estimators of the gender probability ratio of getting a given job



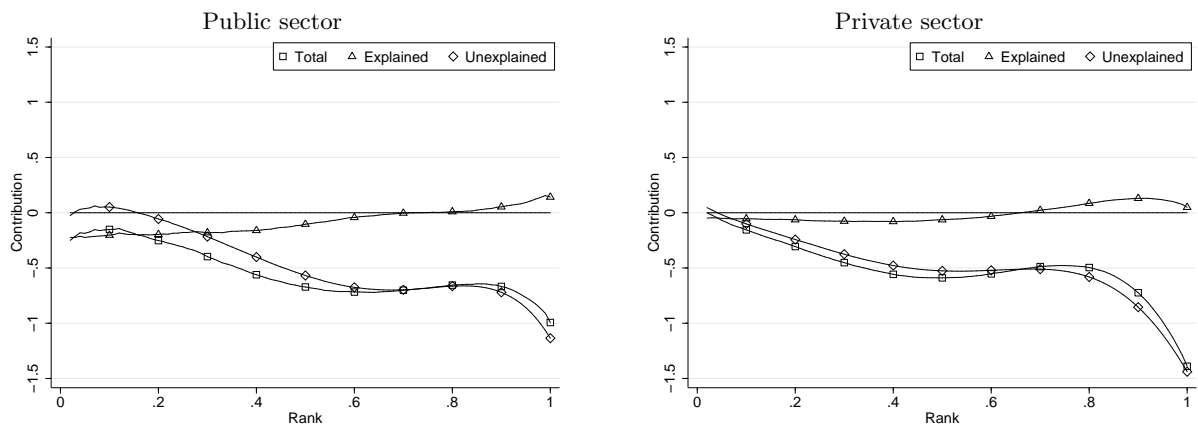
Note: The non-parametric estimator is obtained by applying the empirical strategy proposed by Gobillon, Meurs and Roux (2015). The semi-parametric estimator is obtained by applying the empirical strategy proposed in the current paper. The confidence interval at the 5% level obtained by bootstrap using 100 replications is reported in dotted lines.

Figure 5: Exponentiated gender difference in the effects of dummy variables on the gender probability ratio of getting a given job



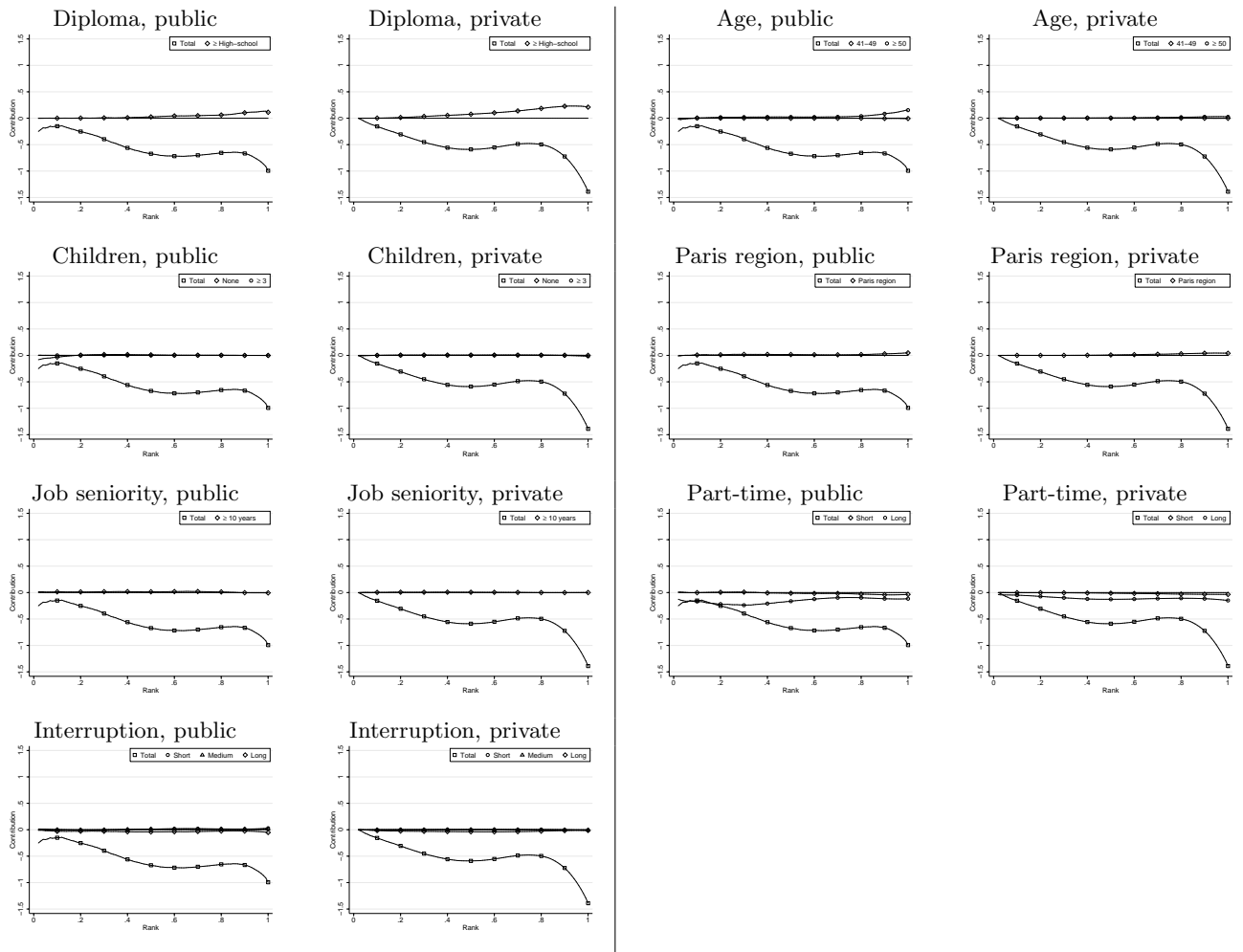
Note: The graph title gives the characteristic for which the exponentiated gender difference in the effects of dummy variables are graphed, as well as the sector. The category corresponding to the reference is mentioned since the corresponding curve is the same across graphs for a given sector. Indeed, this curve represents the gender probability ratio of getting a job as a function of rank for a worker with the values of all the observable characteristics fixed to the reference.

Figure 6: Oaxaca decomposition of the gender probability ratio of getting a given job



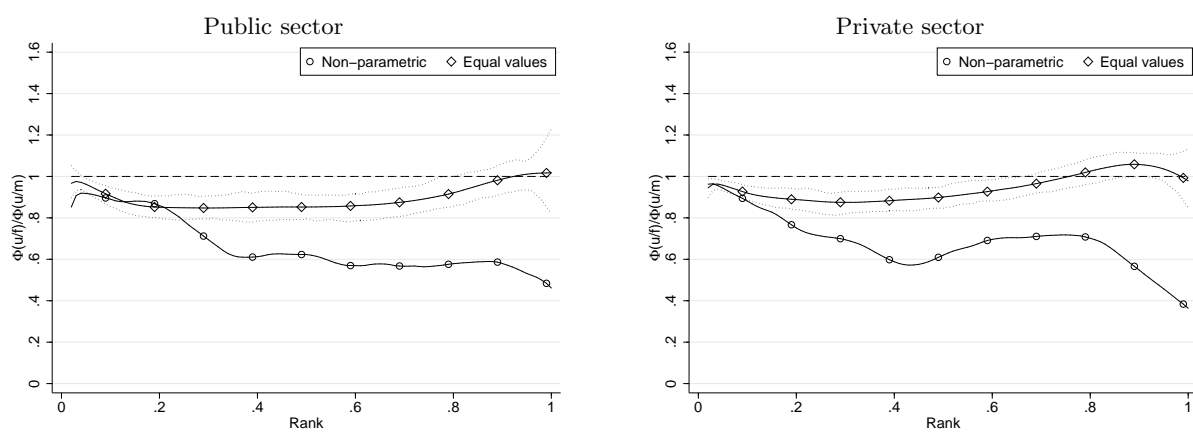
Note: “Total”: gender difference in the logarithm of the average probability of getting a given job; “Explained”: part of “Total” that can be attributed to the gender difference in observable characteristics valued using the estimated coefficients obtained for the whole population; “Unexplained”: part of “Total” that can be attributed to the deviation of gender coefficients of observable characteristics from the ones of the whole population. Note that “Total” is not exactly equal to the sum of “Explained” and “Unexplained” since it also involves a residual term due to the non-linearity of the logarithm function.

Figure 7: Decomposition of the explained part of the gender probability ratio of getting a given job into the contribution of each variable



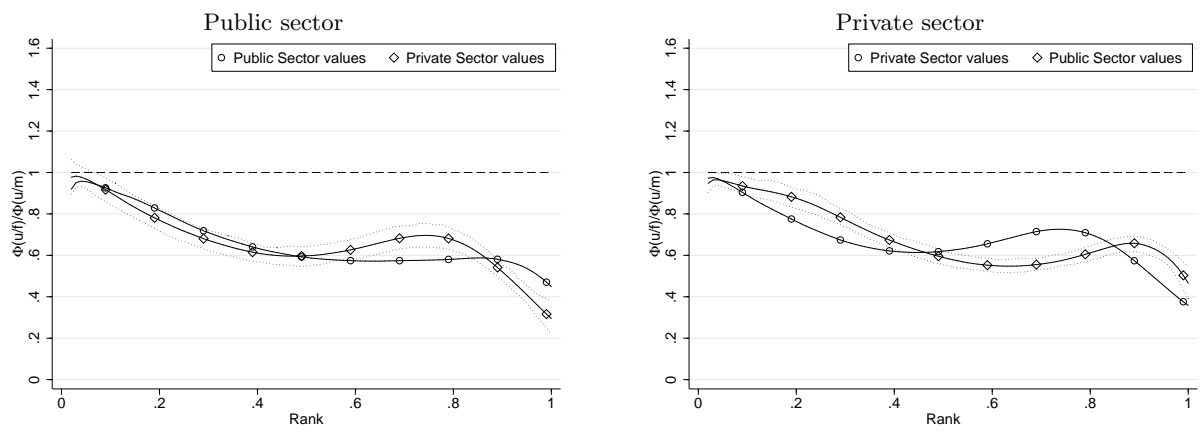
Note: The graph title gives the characteristic for which the exponentiated gender difference in the effects of dummy variables are graphed, as well as the sector. “Total”: gender difference in the logarithm of the average probability of getting a given job; other labels in the legend refer to the category of the variable for which the contribution is graphed.

Figure 8: Counterfactual gender probability ratio of getting a given job when worker conditional values are equal for the two genders



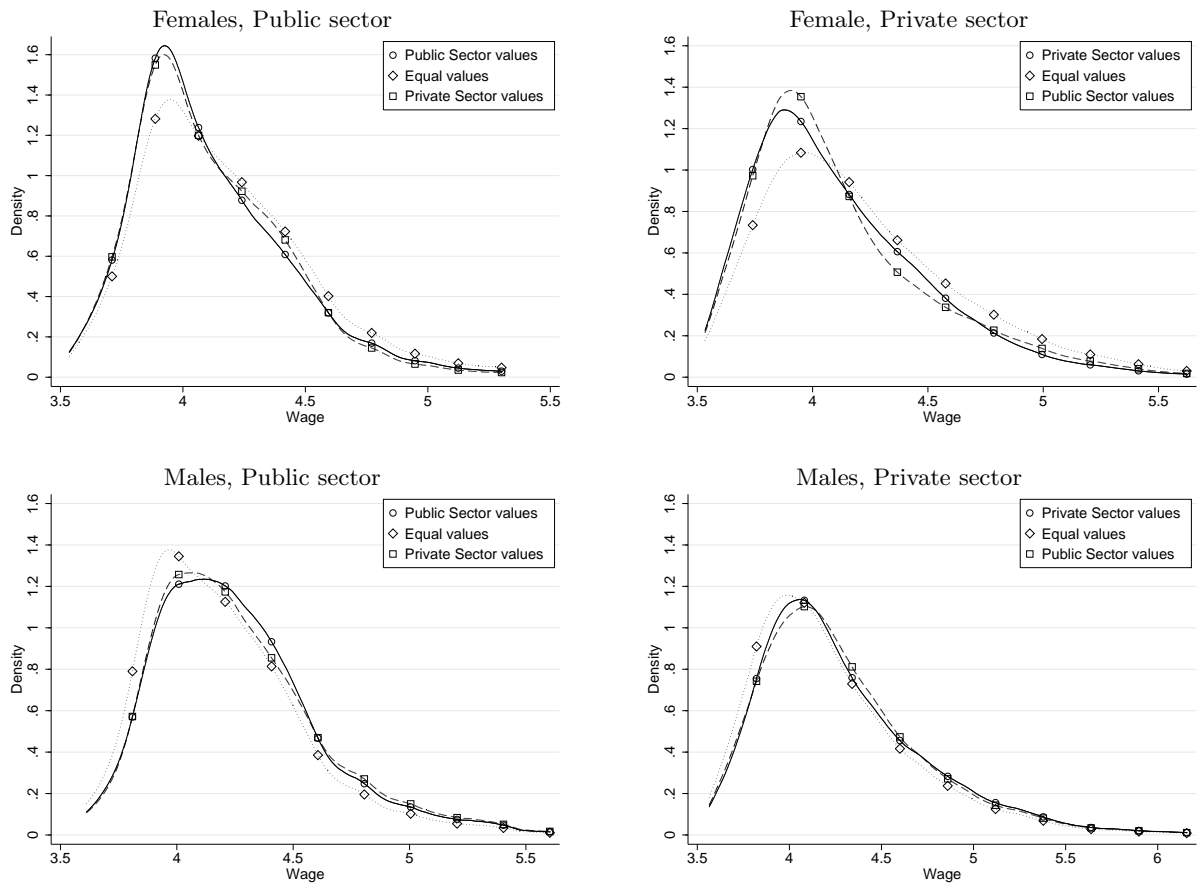
Note: The non-parametric estimator is obtained by applying the empirical strategy proposed by Gobillon, Meurs and Roux (2015). The curve corresponding to “Equal values” is obtained by applying the empirical strategy proposed in the current paper when fixing, for the two genders, the conditional worker values to the same values which are obtained by estimating parameters for the whole population. The confidence interval at the 5% level obtained by bootstrap using 100 replications is reported in dotted lines.

Figure 9: Counterfactual gender probability ratio of getting a given job when conditional worker values are those in the other sector



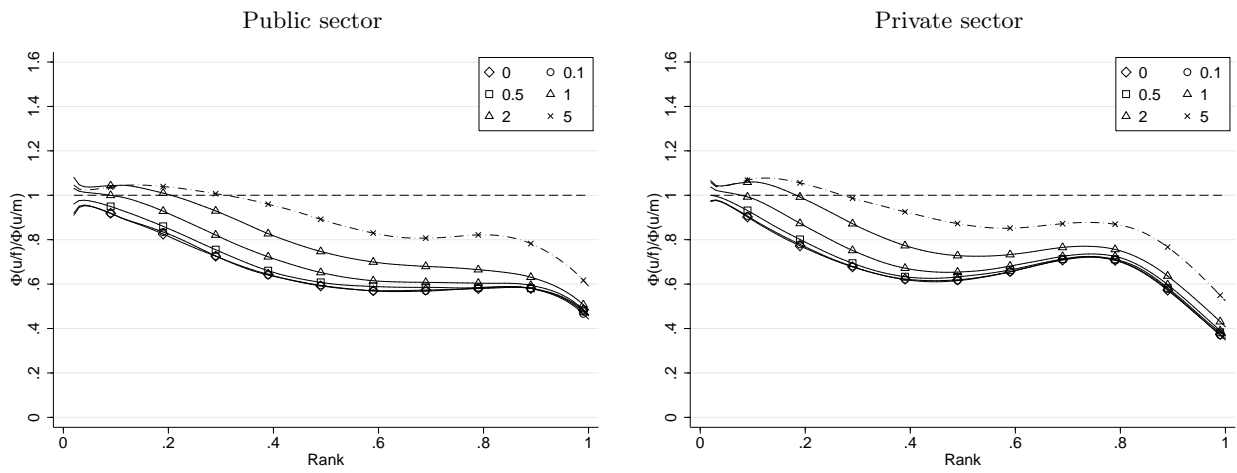
Note: The curve corresponding to “Private Sector values” (resp. “Public Sector values”) is obtained by fixing conditional worker values to those of the private (resp. public) sector. The confidence interval at the 5% level obtained by bootstrap using 100 replications is reported in dotted lines.

Figure 10: Counterfactual log-wage densities in different scenarios



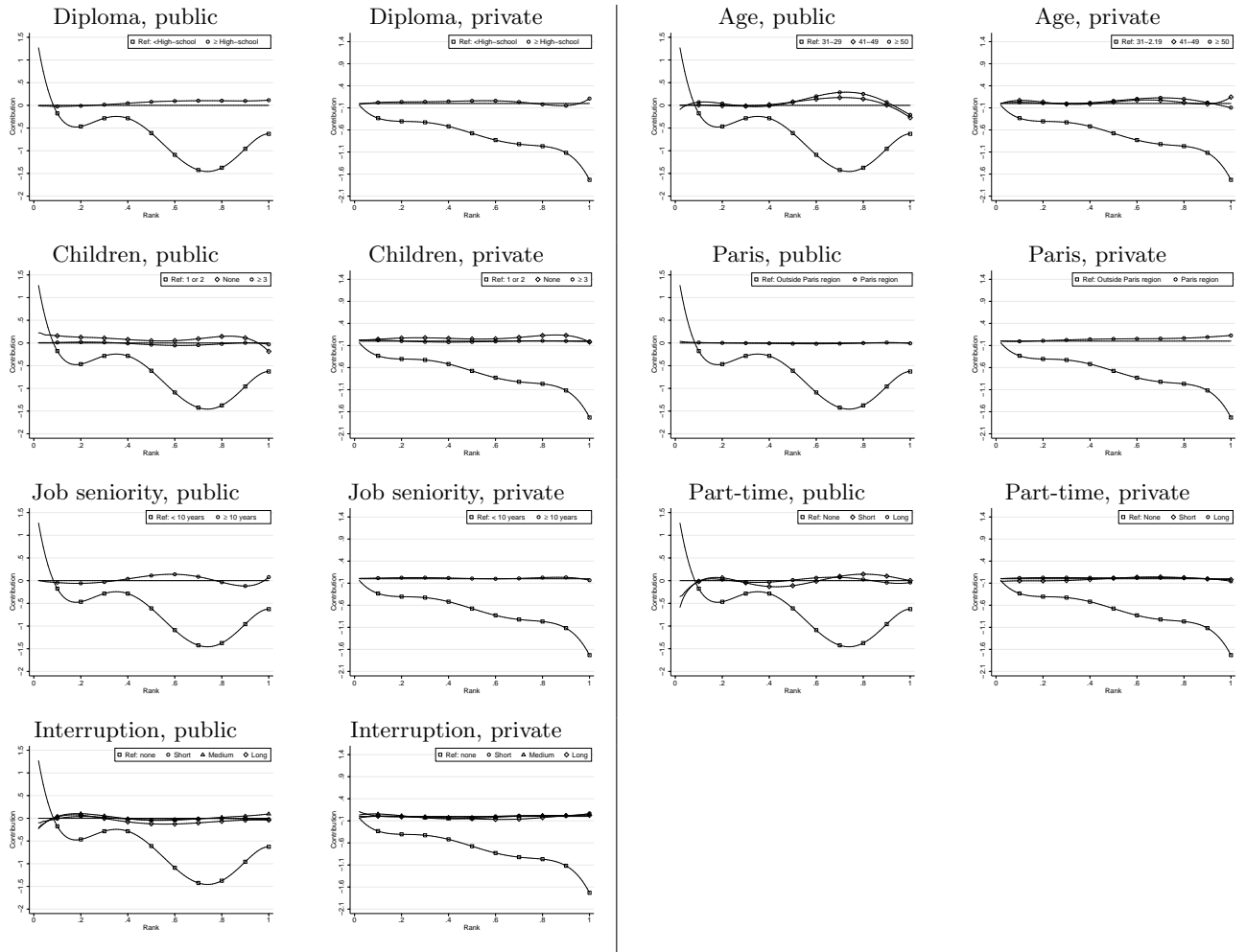
Note: Densities are computed using the logarithm of daily wages generated by the model when reassigning workers to job positions. “Private Sector values”: reassignment using conditional worker values computed for the private sector; “Public Sector values”: reassignment using conditional worker values computed for the public sector; “Equal values”: reassignment using conditional worker values common to the two genders which are obtained by estimating parameters for the whole population.

Figure A.1: Counterfactual gender probability ratio of getting a given job when unobserved individual heterogeneity is simulated



Note: We recompute the gender probability ratio of getting a given job using the estimated parameters and adding unobserved individual heterogeneity terms to the set of individual characteristics. These terms are drawn in a centered normal law with variance equal to the gender-specific variance of the effect of observed individual characteristics multiplied by a parameter which is made to vary and is reported in the graph (see text for details). The confidence interval at the 5% level obtained by bootstrap using 100 replications is reported in dotted lines.

Figure A.2: Decomposition of the unexplained part into the contribution of each variable



Note: The graph title gives the characteristic for which the exponentiated gender difference in the effects of dummy variables are graphed, as well as the sector. Labels in the legend refer to the categories of the variable for which the contribution is graphed.