Wage differentials between native, immigrant and cross-border workers: Evidence and model comparisons

Philippe VAN KERM¹
Seunghoe YU¹,²
Chung CHOE³

CEPS/INSTEAD, Luxembourg¹
KU Leuven, Belgium²
Korea Information Society Development Institute, Korea³
Wage differentials between native, immigrant and cross-border workers: Evidence and model comparisons

Philippe Van Kerm†
CEPS/INSTEAD, Luxembourg

Seunghee Yu
KU Leuven, Belgium and CEPS/INSTEAD, Luxembourg

Chung Choe‡
Korea Information Society Development Institute

March 2014

Abstract

This paper exploits a parametric variant of the Machado-Mata simulation methodology to examine wage distribution differences between native and foreign workers in Luxembourg. Relying on ‘parametric quantile regression’ in place of repeated linear quantile regressions cuts computing time drastically with no loss in the accuracy of unconditional quantile simulations. Substantively, we find a clear inverted-U-shaped native worker advantage: the advantage is small (possibly negative) for both low and high quantiles, but it is large for the middle half of the quantile range (between the 20th and 70th native wage percentiles). The pattern holds against both immigrants and cross-border workers, although the latter catch up much less at high percentiles. Differences in human capital and job characteristics hardly account for the gap, unlike sorting into different jobs and occupations which account for a substantial share—although not all—of the gap.

Keywords: immigrant wages; cross-border workers; quantile regression; quantile process; distribution regression; Singh-Maddala distribution; Dagum distribution; Luxembourg

JEL Classification: J15; J31; J61

---

*This research is part of the InWIn project (Information and wage inequality: Evidence on wage differences between natives, immigrants and cross-border workers in Luxembourg) supported by the Luxembourg ‘Fonds National de la Recherche’ (contract C10/LM/78567). The 2006 Luxembourg Structure of Earnings Survey data were provided by STATEC—Institut national de la statistique et des études économiques. Comments by Stephen Jenkins are gratefully acknowledged.

†Philippe Van Kerm, CEPS/INSTEAD, 3 Av. de la Fonte, L-4364 Esch/Alzette, Luxembourg. E-mail: philippe.vankerm@ceps.lu.

‡Chung Choe (corresponding author), Korea Information Society Development Institute, 36 Jang gun maeu 3 gil, Gwacheon-si, Gyeonggi-do, 427-701, South Korea. E-mail: choechung@kisdi.re.kr. Tel.: +82-2-570-4144.
1 Introduction

Foreign workers are a defining component of the Luxembourg labour market. The 2011 statistics from STATEC indicate that immigrants account for 43.2% of all residents—a fraction higher than in any other OECD country. Additionally, Luxembourg attracts numerous non-resident cross-border workers from Belgium, France or Germany, for a total of 44% of national employment in 2011.¹

Analogously to what has been reported across countries—see, e.g., Butcher and DiNardo (2002), Adsera and Chiswick (2007), Aydemir and Skuterud (2008), Elliott and Lindley (2008) for recent studies and Chiswick (1978) for seminal analysis—, previous studies have shown that both cross-border and immigrant workers receive lower wages than Luxembourg natives (see, e.g., Brosius, 2005, STATEC, 2009, Sologon and Van Kerm, 2014). This is somewhat surprising given their integration in the labour market and because cross-border workers often have higher educational achievements than native workers while immigrant workers are composed of an heterogeneous mix including not only low-skilled but also highly qualified workers (Fusco et al., 2014). Most of the evidence available to date has however documented differences in mean pay, although differences between these groups of workers may not satisfactorily be characterized by a ‘uniform’ wage penalty. One may for example anticipate the existence of “glass ceiling” phenomena limiting access of non-resident and/or non-natives to top management positions within firms.² Or heterogeneity in immigrants’ profiles may lead to different degrees of wage disadvantage for low, middle or high wages, as has, for example, been recently documented in Spain and the United Kingdom by Simón et al. (2008) and Hunt (2012) respectively. For these reasons, this paper sets out to examine in further detail the overall wage distribution differences between native, immigrant and cross-border workers, taking advantage of a large representative data on private sector employees, the 2006 Luxembourg Structure of Earnings Survey.

Detailed analysis of wage distribution differences quantile-by-quantile is now popular in empirical analysis of wage differentials; see the survey by Fortin et al. (2011). Nonetheless, to the best of our knowledge, relatively few studies have looked at the ‘nativity wage gap’ in this way; see Butcher and DiNardo (2002), Canal-Domínguez and Rodríguez-Gutiérrez (2008), Chiswick et al. (2008), Simón et al. (2008), Cai and Liu (2012), Hunt (2012), Diop-Christensen and Hildebrand (2014). Several of these studies find that the wage gap declines with the percentile considered: the gap is large at low wage and tends to decline and often turn negative in the upper part of the wage distribution. The magnitude

¹See http://www.statistiques.public.lu/stat/TableViewer/tableView.aspx.

²We note however that it has also been argued that some immigrants hold top managerial positions in Luxembourg (Debroek, 2009, Amétépé and Hartmann-Hirsch, 2011).
of the gap differs however across destination country and immigrant characteristics.

Virtually all studies apply decomposition analysis to factorize differences in the quantiles of the wage distributions of native-born and foreign workers into ‘price’ and ‘composition’ components and focus discussion on ‘unexplained’ differentials due to differences in labour ‘price’ not immediately justified by human capital and other observable productive characteristics or by occupations and job attributes. This decomposition principle is an immediate extension of the Oaxaca-Blinder decomposition of mean wage differences (Oaxaca, 1973, Blinder, 1973). Application to the ‘quantile wage gap’ is somewhat more complicated since one cannot invoke the law of iterated expectations as in Oaxaca-Blinder equations. A popular solution is the quantile regression-based simulation methodology proposed by Machado and Mata (2005) and later extended in Autor et al. (2005) and Melly (2005). The method—henceforth MMAM05—relies on estimation of a large number of quantile regression models combined with Monte Carlo integration over covariate distributions to construct counterfactual quantiles ‘as if’ foreign workers were rewarded as natives (or vice versa). Such counterfactual quantiles are in turn used to decompose the observed quantile differences into ‘price’ (as per quantile regression coefficients) and ‘composition’ (due to covariates) terms.

MMAM05 requires estimation of a large number of quantile regressions and is therefore computationally intensive. Alternative to MMAM05 is sample reweighting as in Di Nardo et al. (1996) or Barsky et al. (2002) which involves reweighting one of the two samples so as to eliminate differences in observable population characteristics and thereby controlling for composition differences. Reweighting is easier computationally but, since it does not involve estimation of wage equations, it does not allow detailed analysis of the characteristics differentially rewarded in the groups compared or more sophisticated decomposition such as advocated in Juhn et al. (1993) or Melly (2005) for disentangling level from dispersion (or residuals) effects. Our analysis therefore relies on the MMAM05 decomposition but we propose and assess in this paper the performance of a computationally cheaper implementation. Our approach builds upon Biewen and Jenkins (2005) and Noufaily and Jones (2013) and substitutes the computationally intensive quantile regression process estimation in MMAM05 by a ‘parametric quantile regression’ model. This leads to reduction of calculation time by more than ninety percent. Besides the significant gains in computing time, the approach is also flexible in that it allows immediate estimation of (counterfactual) quantiles, means, inequality measures, or distribution functions at once.

Lately, Firpo et al. (2009) have proposed an ‘unconditional quantile regression’ approach which addresses the disadvantage of the reweighting strategy. It relies however on local approximations which may not be accurate if the covariate distributions or the conditional quantiles of the groups compared are not sufficiently close to each other (Rothe, 2012). See Cai and Liu (2012) for an application to Australia and Choe and Van Kerm (2014) for an illustration on the data used in this paper.
and, at least in principle, allows estimation of arbitrarily high or low quantiles. The price to pay is imposition of potentially restrictive parametric assumptions on the (conditional) wage distributions. We therefore examine carefully the fit of alternative models and find that the likelihood-based parametric models fit our data at least as well as the MMAM05 series of quantile regressions (when covariates are included in the model).

On the substantive side, we find an inverted-U-shaped quantile wage gap: the (unexplained) wage gap is negligible at low percentiles, it peaks between the 20th and 70th percentile and then declines at high wages to become nil or even negative, in particular for immigrant workers. Note that this pattern is somewhat different from the downward-sloping gap found in other country studies mentioned above. Differences in labour force composition (defined by sex, education and (potential) experience) explain little of the lower wages paid to foreign workers, at whatever percentile one looks at, but the wage gap is more muted once we account for differences in occupation and job characteristics. It remains however that differences in prices—the hourly wage paid conditional on workers’ human capital for given occupations and sectors—still lead to a significant advantage for native workers, in particular for workers in the middle of the wage distribution. Cross-border and immigrant workers’ wages are very similar in level but immigrant wages a more dispersed and in particular catch up to native wages at the top—a phenomenon that is not observed among cross-border workers.

In the next section, we describe the quantile decomposition that we calculate in the paper and propose a variant to what has been adopted in recent research. We explain our parametric quantile regression strategy in Section 3. Section 4 describes the data and sample definition. Our empirical results are presented in two sections: we thoroughly compare the fit of the parametric quantile estimators with MMAM05 estimators based on quantile regression in Section 5 and we present our substantive decomposition results in Section 6. Section 7 summarizes and concludes.

2 Decomposing the quantile wage gap: composition and price components

In line with much of the literature on quantile wage gaps, we adopt the quantile function decomposition technique proposed by Machado and Mata (2005). The decomposition partitions the observed difference in a given quantile of the wage distributions of two groups of workers into a ‘composition’ (or ‘explained’) component and a ‘price’ (or ‘unexplained’) component; see, e.g., Albrecht et al. (2003), Arulampalam et al. (2007), Antonczyk et al. (2010), Christofides et al. (2013) for recent applications.

Start with the ‘quantile wage gap’ between immigrant and native workers at quantile $\tau$:

$$
\Delta_{\text{Nat}, \text{Immig}}(\tau) = \ln Q_{\text{Nat}}(\tau) - \ln Q_{\text{Immig}}(\tau)
$$

(1)
where $Q_{\text{Nat}}$ and $Q_{\text{Immig}}$ are quantile functions of the wage distributions of native and immigrant workers. To separate out ‘unexplained’ differences between these quantiles from differences that can be accounted for by differences in human capital and job characteristics—the ‘composition’ effect—, define the counterfactual quantile $Q_{j|k}(\tau)$ that would be observed if individuals of group $k$ (natives) were paid as individuals in group $j$ (immigrants). $Q_{j|k}$ is the quantile function of the counterfactual distribution obtained by integrating the conditional wage distribution for group $j$ workers over the group $k$ covariate distribution (Chernozhukov et al., 2013):

$$Q_{j|k}(\tau) = F_{j|k}^{-1}(\tau) \quad (4)$$

and

$$F_{j|k}(y) = \int_{\chi} F_{X(j)|x}(y|x) dH_k(x) \quad (5)$$

where $\chi$ is the domain of definition of characteristics $X$, $F_{X(j)|x}(y|x)$ is the distribution of wages in group $j$ conditional on characteristics $X = x$ and $H_k(x)$ is the probability distribution of covariates $X$ in group $k$. (Notice that $Q_{j}(\tau) = Q_{j|j}(\tau)$.)

The total quantile difference is then decomposed into two components:

$$\Delta_{\text{Nat,Immig}}(\tau) = \left[ \ln Q_{\text{Nat}}(\tau) - \ln Q_{\text{Immig|Nat}}(\tau) \right] + \left[ \ln Q_{\text{Immig|Nat}}(\tau) - \ln Q_{\text{Immig}}(\tau) \right] \quad (6)$$

where the $\Delta_{\text{Immig|Nat,Immig|Immig}}(\tau)$ component captures the ‘composition’ part of the $\tau$-quantile difference which can be accounted for by differences in the characteristics of native and immigrant workers, while the $\Delta_{\text{Nat|Nat,Immig|Immig}}(\tau)$ term is the difference that is left unexplained by differences in workers characteristics alone and that reflects ‘price’ differences embedded in conditional wage distributions differences.

Following Juhn et al. (1993) and Melly (2005), we further factorize the contribution of the price component into ‘level’ and ‘dispersion’ factors (or ‘coefficients’ and ‘residuals’ factors in Melly (2005) terms) to separate out differences in levels of mean pay from differences across groups in higher moments of the respective pay distributions. To do so, define a second counterfactual $\tau$-quantile that

4Our notation throughout this section is inspired by, but differs slightly from, Chernozhukov et al. (2013). We will also examine differences with cross-border workers wages

$$\Delta_{\text{Nat,CB}}(\tau) = \ln Q_{\text{Nat}}(\tau) - \ln Q_{\text{CB}}(\tau) \quad (2)$$

$$\Delta_{\text{Immig,CB}}(\tau) = \ln Q_{\text{Immig}}(\tau) - \ln Q_{\text{CB}}(\tau) \quad (3)$$

but focus in this section on the natives vs. immigrants gap. The other groups will be compared in the same way.
represents the (unconditional) quantile that would be observed in group \( k \) if its workers had the same average pay as workers in group \( j \) (conditionally on their characteristics) but otherwise kept the original relative dispersion of their wage structure:

\[
\hat{Q}_{(j|k)}(\tau) = \tilde{F}^{-1}_{(j|k)}(\tau)
\]

with

\[
\tilde{F}_{(j|k)}(y) = \int_x \tilde{F}_{X_{(j,k)}}(y|x) \, dH_k(x).
\]

\( \tilde{F}_{X_{(j,k)}}(y|x) \) is a rescaled version of \( F_{X_{(j)}}(y|x) \) such that

\[
\tilde{F}^{-1}_{X_{(j,k)}}(y|x) = \frac{\mu_{x(k)}}{\mu_{x(j)}} F^{-1}_{X_{(j)}}(y|x)
\]

(\( \mu_{x(k)} \) and \( \mu_{x(j)} \) are mean wages in groups \( k \) and \( j \) conditional on \( X = x \). \( \tilde{F}_{X_{(j,k)}}(y|x) \) has the mean of \( F_{X_{(j)}}(y|x) \) but the Lorenz ordinates of \( F_{X_{(k)}}^{-1}(y|x) \); it is obtained by inflating or deflating all group \( k \) quantiles by a ratio to match the means of group \( j \) (all conditional on \( X = x \)). Such a transformation of the distribution function is routinely applied in mean-inequality decompositions to separate out differences in mean (holding the Lorenz curve—and therefore relative inequality—constant) from relative inequality changes (holding mean wage constant) (see, e.g., Datt and Ravallion, 1992, Jenkins and Van Kerm, 2005).

We use \( \hat{Q}_{(j|k)}(\tau) \) to further split the ‘price’ component of Eq. (6) into a ‘level’ term and a ‘dispersion’ term as before:

\[
\Delta_{\{\text{Nat|Nat,Immig|Nat}\}}(\tau) = \ln\hat{Q}_{\{\text{Nat}\}}(\tau) - \ln\hat{Q}_{\{\text{Immig|Nat}\}}(\tau) = \left[\ln\hat{Q}_{\{\text{Nat}\}}(\tau) - \ln\tilde{Q}_{\{\text{Immig|Nat}\}}(\tau)\right] \quad \text{‘level’}
\]

\[
+ \left[\ln\tilde{Q}_{\{\text{Immig|Nat}\}}(\tau) - \ln\hat{Q}_{\{\text{Immig|Nat}\}}(\tau)\right] \quad \text{‘dispersion’}
\]

The ‘level’ component captures the contribution of aggregate differences in the level of prices paid for workers of different nationality to the overall wage gap while the ‘dispersion’ captures the residual differences due to differences in the dispersion of wages around their mean.

Note that while the decomposition is analogous to those developed in Juhn et al. (1993) and Melly (2005), we adopt a different labelling of terms. The ‘level’ component is referred to as the ‘coefficient’ effect in Melly (2005) and the ‘dispersion’ and ‘composition’ factors are referred to as respectively ‘residual’ and ‘characteristics’. Melly’s (2005) typology originates in analogy to the OLS framework used in Juhn et al. (1993) which involved only comparisons across groups of (OLS) regression coefficients for mean wages and of the dispersion of OLS residuals. Although analogous in interpretation and
purpose, our decomposition is mathematically different from both Juhn et al. (1993) and Melly (2005). Like Melly’s (2005) approach and all recent research, we relax the homoscedasticity assumption inherent to the Juhn et al. (1993) coefficient-and-residual swapping procedure by estimating conditional quantile functions (see supra). But the decomposition set out in Eq. (10) also deviates from Melly (2005) in that we separate out ‘level’ and ‘dispersion’ by rescaling mean wages and Lorenz curves, where Melly (2005) takes the median as ‘level’ and is then left with a residual distribution around the median that has no immediate interpretation (unlike the Lorenz curve).

3 Estimation

Evaluation of the decompositions set out in (6) and (10) require estimation of the counterfactual quantiles \( Q_{\langle j \mid k \rangle}(\tau) \) and \( \tilde{Q}_{\langle j \mid k \rangle}(\tau) \). MMAM05 show how these can be conveniently simulated from conditional quantile estimates based on quantile regression.

3.1 Simulating counterfactual quantiles

\( F_{\langle j \mid k \rangle} \) can be understood as the distribution function of a counterfactual outcome created by first sampling \( x \) from the covariate distribution \( H_k \) and then sampling the outcome \( y \) from the conditional wage distribution \( F_{X\langle j \rangle}(\cdot \mid x) \) (Chernozhukov et al., 2013). Building upon this representation, Machado and Mata (2005) propose a Monte Carlo simulation algorithm to generate simulated samples from \( F_{\langle j \mid k \rangle} \) and use the empirical quantiles from the simulated samples to calculate the decomposition components in Eq. (6) above. In a first step, a random draw of \( x_d \) is taken from the observed distribution of covariates in group \( k \) to simulate a draw from \( H_k \). In a second step, inverse sampling is adopted to draw from \( F_{X\langle j \rangle}(\cdot \mid x_d) \): \( u \) is randomly drawn from a uniform distribution \( U(0,1) \) and \( y = F_{X\langle j \rangle}^{-1}(u \mid x_d) = Q_{X\langle j \rangle}(u \mid x_d) \) is taken as a realization from \( F_{\langle j \mid k \rangle} \). These two steps are repeated a large number of times to generate a simulated sample from the counterfactual distribution.

Practically, Machado and Mata (2005) specify a linear quantile regression model for \( Q_{X\langle j \rangle}(u \mid x_d) \):

\[
Q_{X\langle j \rangle}(u \mid x_d) = P(x_d)\beta_j(u) \tag{11}
\]

where \( P(x_d) \) is a vector of (transformations of) \( x_d \) and the quantile regression coefficients \( \beta(u) \) are solutions to the following minimization on the sample of observations from population \( j \)

\[
\hat{\beta}(u) = \arg \min_{\beta} \sum_{i=1}^{n_j} \rho_u(y_i - P(x_i)\beta_j(u)) \tag{12}
\]

with check function \( \rho_u(s) = s(u - 1(s < 0)) \) (Koenker and Bassett, 1978, Koenker and Hallock, 2001).
Autor et al. (2005) and Melly (2005) propose variations of this algorithm. The modified algorithms essentially (i) replace the random sampling of $x$ in the first stage by taking the whole sample from population $k$ and (ii) replace random sampling from $U(0, 1)$ in the second stage by systematic estimation of the full quantile process $Q_{X(j)}(u|x)$ using a fine mesh of $K$ values for $u \in (\epsilon, 1 - \epsilon)$ (where $\epsilon$ is set as small as possible to capture tails of the conditional distributions). This procedure generates a simulated pseudo-sample of size $n_k \times K$ drawn from the counterfactual $F_{(j|k)}$ from which empirical quantiles can be calculated as in the Machado and Mata (2005) algorithm to obtain $Q_{(j|k)}(\tau)$. Albrecht et al. (2009) and Chernozhukov et al. (2013) demonstrate that these procedures yield consistent estimates of the counterfactual quantiles $Q_{(j|k)}(\tau)$ and are therefore valid ways to estimate the components of (6).

$\tilde{Q}_{(j|k)}(\tau)$ is similarly estimated by rescaling all conditional quantile estimates by the ratio of conditional means of the two groups compared prior to generating the simulated pseudo-sample by stacking the conditional quantile vectors across the samples covariate and calculating unconditional quantiles.  

3.2 Quantile regression estimation vs. parametric distribution models

The simulation algorithms just described provide simple ways to estimate the counterfactual quantiles needed to disentangle ‘price’ and ‘composition’ components of quantile differences. By specifying $P(X)$ flexibly (e.g., considering a fully saturated model in case of discrete covariates), setting $\epsilon$ close to zero and choosing $K >> 0$, one can achieve any desired level of accuracy in the approximation of $Q_{(j|k)}(\tau)$ (Chernozhukov et al., 2013).

In practice, however, limited sample sizes impose constraints on the flexibility in $P(X)$ which can be afforded and on minimal acceptable values for $\epsilon$. Furthermore, the computational cost of estimating a large number of quantile regressions and of estimating empirical quantiles in a potentially very large ($n_k \times K$) vector of simulated data imposes practical limits on $K$. Computational burden is compounded with bootstrap inference to assess sampling variability (as recommended in Chernozhukov et al., 2013).

We consider a computationally cheaper alternative. This involves parameterizing a model for the conditional distribution $F_{X(j)}(\cdot|x_d)$ and deriving the conditional quantile process from a small set of model parameters. This is akin to relying on what Noufaily and Jones (2013) coin a ‘parametric quantile regression’ instead of the classic quantile regression in the second stage of the MMAM05 algorithm (or its variants).

The models for the conditional distributions that we consider are the log-normal (Aitchison and

---

5Conditional means can be estimated from a separate model for the conditional mean—e.g., regressions—or more simply by averaging conditional quantile estimates.
Brown, 1969), the Singh-Maddala (Singh and Maddala, 1976) and the Dagum (Dagum, 1977) distributions. All three are skewed, unimodal distributions frequently adopted in modelling income distributions. Kleiber and Kotz (2003) provide a thorough discussion of the properties of these three distributions. McDonald (1984) shows the relationships between the Dagum and Singh-Maddala models (both are members of the broader class of Generalized Beta distributions of the second kind).

The quantile functions of all three distributions have simple closed-form expressions and can therefore be recovered easily from parameter estimates. Given parameters estimates \( m \) and \( s \), the quantile function for the log-normal distribution is given by

\[
Q_{LN}(u; a, b, q) = \exp \left( m + s \Phi^{-1}(u) \right),
\]

(13)
given parameters estimates \( a, b \) and \( q \), the quantile function for the Singh-Maddala distribution is given by

\[
Q_{SM}(u; a, b, q) = b \left( (1 - u)^{-\frac{1}{q}} - 1 \right)^{\frac{1}{a}},
\]

(14)
and for the Dagum distribution with parameters \( a, b \) and \( p \) by

\[
Q_{D}(u; a, b, p) = b \left( u^{-\frac{1}{p}} - 1 \right)^{-\frac{1}{a}}.
\]

(15)
(Kleiber and Kotz, 2003). Following Biewen and Jenkins (2005), we condition these distributions on covariate vectors \( X \) by letting each of the two or three parameters vary linearly with \( P(X) \), that is, for any parameter \( \theta_x \), we let

\[
\theta_x = P(x)\beta_\theta.
\]

(16)
Maximum likelihood estimation of all \( \beta_\theta \) parameters is straightforward using standard Newton-Raphson optimization.\(^6\)

The parametric specifications are obviously more restrictive than the linear quantile regression approach: the latter effectively uses \( K \times g \) parameters to model the conditional distributions (\( g \) is the dimension of \( P(X) \)) while our parametric assumptions involve only \( p \times g \) parameters where \( p = 2 \) or \( p = 3 \). However, imposition of a parametric model results in tremendous savings on computation times since only one likelihood function need to be maximized instead of \( K \) separate quantile regressions. The price to pay is the reliance on parametric modelling assumptions whose validity must be empirically assessed. When computation time is an issue, provided the parametric assumptions are tenable, this strategy is a much more satisfactory way to save computation time than, say, relying on high \( \epsilon \) or small \( K \) in predicting the full quantile process by quantile regression. Setting \( \epsilon \) high essentially

\(^6\)We estimated all models in Stata (StataCorp, 2011) and used the user-written routines \texttt{lognfit}, \texttt{dagumfit} and \texttt{smfit} provided by Jenkins (2004).
trims the simulated data from extreme draws from the conditional distributions and therefore potentially underestimates both very high and very low incomes in the simulated sample; reducing the size of $K$ results in simulation based on a coarse set of prediction of the conditional quantile process. Such compromises can be avoided by relying on parameterized conditional distribution estimates.

Note finally that restrictions in modelling $P(X)$—such as additivity in covariates—may potentially lead to specification error in counterfactual sampling from $F_{(j|k)}(\cdot)$ based on quantile regression too. Linearity of quantiles with respect to $P(x)$ as per Eq. (11) is different from the linearity of model parameters assumed in Eq. (16). Since the two approaches are not nested, there is no guarantee that the linear quantile regression will necessarily result in more accurate sampling from $F_{(j|k)}$ than the ‘parametric quantile regression’, despite the higher number of free parameters. Results will show that, in fact, parametric models often lead to more accurate predictions of the observed quantile functions in our data than 99 quantile regression processes (see Section 5).

3.3 Fit comparisons

Whether simulations based on ‘parametric quantile regression’ result in satisfactory approximations of the unconditional distributions and how they compare with the standard linear quantile regression approach is essentially an empirical question.

We first assess and compare the fit of the estimated conditional quantiles from the value of the average of the check function $\rho_\tau(\cdot)$ (Landejo et al., 2008):

$$E_j \rho(\tau) = \frac{1}{n_j} \sum_{i=1}^{n_j} (y_i - Q_{X(j)}(\tau | x_i)) (\tau - 1((y_i - Q_{X(j)}(\tau | x_i)) < 0))$$

computed with values of $Q_{X(j)}(\tau | x_i)$ predicted by the different models, that is from equations (11), (13), (14) or (15). The value of the check function is the quantile regression equivalent to the squared residuals in OLS models. The model returning the smallest expected value thereof is interpreted as having a better fit to the observed data. Note that the measure $E_j \rho(\tau)$ is ‘local’ to a particular conditional quantile $\tau$.

Secondly, instead of considering the fit of the conditional quantile functions directly, we inspect the fit of the model-based simulated approximations to $Q_{(j)}(\tau) = Q_{(j|j)}(\tau)$ with the direct empirical estimates on these unconditional distribution quantiles (for $j$ being each of native, immigrant and cross-.

---

7As is standard, our analysis will rely on additive specifications for $P(X)$. Most of our covariates are binary or categorical and linear splines are used for the two continuous covariates (namely age and job tenure); see Section 4 and Appendix A for details. The effect of each covariate is therefore modelled flexibly, however we assume additivity of covariates and only limited interaction effects are introduced in $P(X)$. Non-parametric, kernel-based enhancements would not reveal particularly attractive as they lead to dramatic increases in computational cost.
border workers). To summarize the fit, we will report the average absolute and squared deviations over 49 quantiles covering P2 to P98 at which unconditional quantiles are predicted. We also assess the distance between the observed and the predicted quantile functions on the basis of the family of indices recently proposed in Cowell et al. (2009). We calculate the $J_\alpha$ family of goodness-of-fit measures

$$J_\alpha = \frac{1}{K \alpha (\alpha - 1)} \sum_{t=1}^{K} \left[ \left( \frac{Q_j(t)}{\mu_j} \right)^{\alpha} \left( \frac{Q_{\langle j \rangle}(t)}{\mu_{Y,j}} \right)^{1-\alpha} - 1 \right]$$

(18)

where $Q_j(t)$ is the $t$-th empirical quantile of the observed wage distribution and $Q_{\langle j \rangle}(t)$ is the corresponding quantile simulated from the models, and $\mu_j$ and $\mu_{Y,j}$ are the averages of these quantiles. Varying the $\alpha$ parameter changes the weight given to different parts of the wage distribution: the higher is $\alpha$ the greater is the sensitivity to deviations in the top end of the distribution. We consider $\alpha$ between $-1$ (bottom sensitive) and $+1$ (top sensitive). Note that this definition of $J_\alpha$ is an approximation of the measure given in Cowell et al. (2009): to keep estimation tractable, we consider the distance between the observed and simulated curves only at the $K$ grid point, not on the full sample. While reducing computation time, this approximation also bounds the leverage on $J_\alpha$ of extreme observations with wages above and below the 2nd and 98th percentiles of the wage distributions.

4 Data

We use data from the 2006 Luxembourg Structure of Earnings Survey (SES). The data were collected in all European Union countries on the basis of common variable definitions and sampling design defined in European Community regulations. It provides detailed information about the structure of earnings in 27 Member States. The Luxembourg SES is collected by STATEC—Institut national de la statistique et des études économiques, the national statistical institute. The distinctive feature of the SES in the context of Luxembourg is that—since it is based on a sampling frame of firms—it collects information on both resident and cross-border workers.

The 2006 SES is designed to capture representative data for non-profit and private sector firms (NACE C–K and M–O) with at least 10 workers on the level of remuneration and individual characteristics of employees (sex, age, occupation, tenure, highest educational level attained, etc.) and their employer (economic activity, size and location of the enterprise). This matched employer-employee survey covers 79 percent of salaried workers in Luxembourg at the time of the survey (STATEC, 2009).\(^8\)

In total, the 2006 Luxembourg SES dataset covers 1,856 firms and 31,329 workers (STATEC, 2009).

\(^8\)Workers in the public sector and agriculture are excluded from the sampling frame of the survey. These sectors employ only few immigrants or cross-border workers.
Our estimation sample is restricted to workers aged 25 to 59 and contains observations of 6,857 Luxembourg nationals, 7,581 immigrant workers (foreign residents), and 13,875 cross-border workers (non-residents). Hourly wage is computed by dividing the monthly wage of October 2006 by the number of paid hours worked. We dropped observations with calculated hourly wage less than 5 euros and observations with hourly wages above 250 euros to limit the influence of the most extreme cases.

In line with earlier evidence in the immigration literature, immigrant and cross-border workers in Luxembourg earn much less than native counterparts. Figure 1 and Table 1 show the wage distributions among native, immigrant and cross-border workers in our sample. Native workers’ wages are substantially higher, irrespectively of the percentile we consider (except only for the smallest percentile reported, namely the 2nd percentile). Mean wage for natives (€24.0) is 23 percent higher than mean wage among immigrants (€18.6) and 24 percent higher than mean wage among cross-border workers (€18.3). The difference is even more striking at the median, with the median wage among natives being 33 and 26 percent higher than the median among immigrants and cross-border workers respectively. Comparison of the two categories of foreign workers reveals that the wage distributions of immigrants and cross-border workers are similar in levels—mean wage is almost identical in the two groups—but the quantile function for cross-border workers is flatter with higher wages than immigrants up to the 75th percentile, and markedly lower wages in the top 10 percent. Note how the top immigrant wages converge towards top native wages; this phenomenon is not observed in cross-border workers’ wages. The comparatively flatter quantile function for cross-border workers is reflected in the lower percentile ratios reported at the bottom of Table 1 compared in particular to immigrant workers. Native wages are undeniably higher than foreign wages, but the gap is not uniform across all wage levels.

<table>
<thead>
<tr>
<th></th>
<th>Native workers</th>
<th>Immigrant workers</th>
<th>Cross-border workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>€ 24.0</td>
<td>€ 18.6</td>
<td>€ 18.3</td>
</tr>
<tr>
<td>10</td>
<td>€ 11.6</td>
<td>€ 9.3</td>
<td>€ 10.4</td>
</tr>
<tr>
<td>20</td>
<td>€ 14.4</td>
<td>€ 10.7</td>
<td>€ 11.7</td>
</tr>
<tr>
<td>50</td>
<td>€ 21.2</td>
<td>€ 14.3</td>
<td>€ 15.3</td>
</tr>
<tr>
<td>80</td>
<td>€ 31.0</td>
<td>€ 23.8</td>
<td>€ 23.2</td>
</tr>
<tr>
<td>90</td>
<td>€ 37.6</td>
<td>€ 32.3</td>
<td>€ 29.4</td>
</tr>
<tr>
<td>P90/P10</td>
<td>3.25</td>
<td>3.47</td>
<td>2.82</td>
</tr>
<tr>
<td>P90/P50</td>
<td>1.77</td>
<td>2.25</td>
<td>1.92</td>
</tr>
<tr>
<td>P10/P50</td>
<td>0.54</td>
<td>0.65</td>
<td>0.68</td>
</tr>
</tbody>
</table>


Estimates reported in Figure 1 and Table 1 are unadjusted for any difference in productive charac-

11
Figure 1: Hourly wages of native, immigrant and cross-border workers in the Luxembourg private sector in 2006: percentiles and mean wages

Note: Estimates reported from the 2nd to the 98th percentiles. Private sector workers aged 25 to 59. Source: Luxembourg Structure of Earnings Survey 2006. See text for details on the sample and wage definition.

teristics of the groups of workers. Table 2 shows sample means for the three groups of workers for the covariates used in our analysis. Relative to foreign workers, native workers are more likely to work in large firms and to be managers, and they are older, having longer job tenure. Foreign workers are more likely to work in the sector of construction, hotel/restaurant, and real estate. It is also noteworthy that the distribution of educational attainments of immigrants is polarized—with a higher fraction of both primary and tertiary education level. Immigrant workers are more likely to hold low skilled jobs, whereas native workers are more likely to hold professional jobs. Cross-border workers generally have higher educational achievements, but are also younger and have the lowest job tenure.
Table 2: Sample means of hourly wage, nationality and human capital and job characteristics by worker types (Luxembourg nationals, immigrants and cross-border workers)

<table>
<thead>
<tr>
<th></th>
<th>Luxembourg nationals</th>
<th>Immigrant workers</th>
<th>Cross-border workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hourly wage</td>
<td>23.99</td>
<td>18.56</td>
<td>18.31</td>
</tr>
<tr>
<td>Female</td>
<td>0.39</td>
<td>0.38</td>
<td>0.32</td>
</tr>
<tr>
<td>Age</td>
<td>41.23</td>
<td>38.87</td>
<td>38.31</td>
</tr>
<tr>
<td>Primary or less</td>
<td>0.12</td>
<td>0.24</td>
<td>0.08</td>
</tr>
<tr>
<td>Secondary (ref)</td>
<td>0.80</td>
<td>0.61</td>
<td>0.80</td>
</tr>
<tr>
<td>Tertiary</td>
<td>0.09</td>
<td>0.14</td>
<td>0.12</td>
</tr>
<tr>
<td>Years at current employer</td>
<td>12.69</td>
<td>6.74</td>
<td>5.98</td>
</tr>
<tr>
<td>Manager</td>
<td>0.18</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>10-49 employees</td>
<td>0.23</td>
<td>0.31</td>
<td>0.27</td>
</tr>
<tr>
<td>50-249 employees</td>
<td>0.23</td>
<td>0.30</td>
<td>0.35</td>
</tr>
<tr>
<td>250-499 employees (ref)</td>
<td>0.11</td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td>500-999 employees</td>
<td>0.09</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>1000 or more employees</td>
<td>0.34</td>
<td>0.14</td>
<td>0.12</td>
</tr>
<tr>
<td>Part time</td>
<td>0.18</td>
<td>0.15</td>
<td>0.13</td>
</tr>
<tr>
<td>Industry/Manufacture</td>
<td>0.17</td>
<td>0.11</td>
<td>0.18</td>
</tr>
<tr>
<td>Construction</td>
<td>0.04</td>
<td>0.21</td>
<td>0.13</td>
</tr>
<tr>
<td>Wholesale</td>
<td>0.11</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>Hotel/Restaurant</td>
<td>0.01</td>
<td>0.06</td>
<td>0.03</td>
</tr>
<tr>
<td>Trans/Comm</td>
<td>0.16</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>Finance</td>
<td>0.19</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>Real estate</td>
<td>0.08</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td>Education, Health and Other not-for-profit (ref)</td>
<td>0.23</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>Managerial</td>
<td>0.08</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>Professional</td>
<td>0.10</td>
<td>0.09</td>
<td>0.12</td>
</tr>
<tr>
<td>Associate professional</td>
<td>0.24</td>
<td>0.13</td>
<td>0.18</td>
</tr>
<tr>
<td>Clerk</td>
<td>0.23</td>
<td>0.11</td>
<td>0.15</td>
</tr>
<tr>
<td>Service worker</td>
<td>0.07</td>
<td>0.09</td>
<td>0.10</td>
</tr>
<tr>
<td>Craft and trade worker</td>
<td>0.12</td>
<td>0.21</td>
<td>0.20</td>
</tr>
<tr>
<td>Manufacturers</td>
<td>0.08</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>Low skilled and laborer (ref)</td>
<td>0.07</td>
<td>0.20</td>
<td>0.07</td>
</tr>
<tr>
<td>Number of observations</td>
<td>6860</td>
<td>7582</td>
<td>13877</td>
</tr>
</tbody>
</table>

Notes: Based on the 2006 Luxembourg Structure of Earnings Survey. Sample weights applied.
5 Model comparisons: parametric versus classical quantile regression

In this section we assess the extent to which the ‘parametric quantile regression’ models based on the log-normal, Singh-Maddala or Dagum distributions provide satisfactory estimates of wage distributions in comparison to the more flexible (but computationally heavier) estimation of the full (conditional) quantile process by quantile regression.

Figure 2 shows first—as per Eq. (17)—the average value of the check function at each of 99 equi-distant conditional quantiles from 0.01 to 0.99 at which predictions from the models were taken. Remember that the quantile regression parameters are defined as the parameters which minimize the check function; see Eq. (12). The average of the check function for a given percentile can be interpreted similarly to the mean squared residuals in OLS regression. The striking observation from Figure 2 is that the quantile predictions based on parametric models (the dotted and dashed lines) often lead to lower average check function values than with quantile regression (solid line). This means that parametric model predictions generally have a better fit to the conditional quantiles than the quantile regression predictions, except for the highest quantiles. This comes as a surprise since estimation of the quantile process by quantile regressions has many more free parameters than the parametric likelihood-based specifications. Remember however that the two approaches are not nested; in the parametric models, the relationship between the model parameters and predicted quantiles is highly non-linear (Noufaily and Jones, 2013).

The log-normal distribution generally has the worst fit of the three parametric models. There is no clear ‘winner’ between the Singh-Maddala and the Dagum specifications, although the Dagum distribution tends to give a better fit than the Singh-Maddala distribution for the immigrants and cross-border workers models.

The second set of goodness-of-fit checks is based on comparison of simulated unconditional quantile functions with the empirical estimates obtained directly from the raw data (as shown in Figure 1). Figure 3 shows the empirical and the simulated quantile functions for the three subgroups of workers and with two alternative sets of covariates according to the four models. Simulations were conducted as described in Section 2 (also see Appendix A for further implementation details).

All estimated quantile functions are reassuringly close to each other. Simulated and observed unconditional quantiles are often visually undistinguishable. In the models including all covariates (including job and occupation characteristics), simulated quantiles deviate by maximum 5 percent from the observed quantiles (except for predictions at the 2nd and 98th percentiles). (See Appendix Figures B.1 and B.2 for relative deviation estimates.) The fit is not as good when only demographic and human
Figure 2: Average check function residual by conditional quantile for different model specifications; demographic and human capital covariates only (top) and with additional occupation and job characteristics (bottom)

**Human capital covariates only:**

(a) Native workers

(b) Immigrant workers

(c) Cross-border workers

**Human capital and job characteristics:**

(d) Native workers

(e) Immigrant workers

(f) Cross-border workers
capital covariates are considered but deviations remain mostly within 10 percent except at the most extreme quantiles. The fit deteriorates mainly at the very top of the distribution: while the parametric models tend to overestimate the very top percentiles, the quantile regressions tend to underestimate them. Note that, again, the quantile regression-based simulations do not reveal superior to the other parametric quantile model predictions. The log-normal distribution however shows noticeably worse performance for immigrant and cross-border workers with the smaller set of covariates.

Tables 3 and 4 summarize the deviations of the different models by averaging the differences between the observed and simulated quantiles over all 49 equi-distant percentiles from $	ext{P2}$ to $	ext{P98}$ at which they were calculated. They confirm the good fit of the models and that quantile regressions do not necessarily provide a better fit to the data than the parametric models, in particular with the larger set of covariates. The tables also report the $J_{\alpha}$ deviation statistics for different values of the $\alpha$ parameter (higher $\alpha$ place higher weight to divergence at high wages). Again, parametric models turn out to provide satisfactory estimates. The $J_{\alpha}$ index is in fact ‘better’ for the quantile regression models only in one instance: the cross-border model with only human capital and demographic controls. In all other 5 instances at least one parametric model outperforms it. However, no clear ‘winner’ emerges: depending on the subgroup and covariate set considered the best fit can be achieved by any of the three parametric models. We note however that the Singh-Maddala and log-normal exhibit markedly inferior fit among immigrant and cross-border workers with only human capital and demographic characteristics, while the Dagum model tends to perform well in all three subgroups. When job characteristics are introduced, the comparative performance of the Dagum model deteriorates, however its overall fit remains satisfactory. On this basis, we will make it our model of choice throughout Section 6 (and report alternative estimates in Appendix D).

Tables 5 and 6 finally show the prediction error on specific distributional statistics, including the unconditional mean and a series of inequality measures. The discrepancy between the model-based simulations and the empirical estimates can be large for inequality measures. This is unsurprising since the latter are more sensitive to the tails of the distributions which are less closely approximated by the models. For example if the prediction error on the Gini coefficient remains below about 10 percent, it shoots up to 56 percent for the GE2 measure which is known to be highly influenced by top incomes (Cowell and Flachaire, 2007). Caution would be required in exploiting the models for detailed inequality analysis.

The various measures of fit examined support application of the ‘parametric quantile’ approach in our data. It results in simulated unconditional quantile estimates no worse—and often better—than the quantile regression process approach despite their much smaller number of free parameters and achieve
Figure 3: Observed and simulated unconditional quantile functions from different model specifications; demographic and human capital covariates only (top) and with additional occupation and job characteristics (bottom)

**Human capital covariates only:**

(a) Native workers
(b) Immigrant workers
(c) Cross-border workers

**Human capital and job characteristics:**

(d) Native workers
(e) Immigrant workers
(f) Cross-border workers
Table 3: Average prediction error statistics and $J_{\alpha}$ deviation indices ($\times 10^4$) (demographic and human capital covariates model)

<table>
<thead>
<tr>
<th></th>
<th>Native workers</th>
<th>Immigrant workers</th>
<th>Cross-border workers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>QR</td>
<td>S-M</td>
<td>Dag</td>
</tr>
<tr>
<td>Average absolute prediction error (euros)</td>
<td>0.29</td>
<td>0.35</td>
<td>0.37</td>
</tr>
<tr>
<td>Root average squared prediction error (euros)</td>
<td>0.82</td>
<td>0.52</td>
<td>0.56</td>
</tr>
<tr>
<td>Average absolute percentage prediction error (percent)</td>
<td>0.95</td>
<td>1.74</td>
<td>1.76</td>
</tr>
<tr>
<td>$J_{\alpha}$ index ($\alpha = -0.5$)</td>
<td>2.714</td>
<td>2.154</td>
<td>2.271</td>
</tr>
<tr>
<td>$J_{\alpha}$ index ($\alpha = 0$)</td>
<td>2.748</td>
<td>2.156</td>
<td>2.261</td>
</tr>
<tr>
<td>$J_{\alpha}$ index ($\alpha = 0.5$)</td>
<td>2.783</td>
<td>2.158</td>
<td>2.250</td>
</tr>
<tr>
<td>$J_{\alpha}$ index ($\alpha = 1$)</td>
<td>2.820</td>
<td>2.160</td>
<td>2.241</td>
</tr>
</tbody>
</table>
Table 4: Average prediction error statistics and $J_\alpha$ deviation indices ($\times 10^4$) (demog., human capital and job char. covariates model)

<table>
<thead>
<tr>
<th></th>
<th>Native workers</th>
<th>Immigrant workers</th>
<th>Cross-border workers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>QR</td>
<td>S-M</td>
<td>Dag</td>
</tr>
<tr>
<td>Average absolute prediction error (euros)</td>
<td>0.51</td>
<td>0.38</td>
<td>0.40</td>
</tr>
<tr>
<td>Root average squared prediction error (euros)</td>
<td>0.65</td>
<td>0.45</td>
<td>0.61</td>
</tr>
<tr>
<td>Average absolute percentage prediction error (percent)</td>
<td>2.54</td>
<td>1.88</td>
<td>1.80</td>
</tr>
<tr>
<td>$J_\alpha$ index ($\alpha = -1$)</td>
<td>3.573</td>
<td>1.942</td>
<td>2.508</td>
</tr>
<tr>
<td>$J_\alpha$ index ($\alpha = -0.5$)</td>
<td>3.591</td>
<td>1.942</td>
<td>2.499</td>
</tr>
<tr>
<td>$J_\alpha$ index ($\alpha = 0$)</td>
<td>3.609</td>
<td>1.941</td>
<td>2.490</td>
</tr>
<tr>
<td>$J_\alpha$ index ($\alpha = 0.5$)</td>
<td>3.626</td>
<td>1.940</td>
<td>2.482</td>
</tr>
<tr>
<td>$J_\alpha$ index ($\alpha = 1$)</td>
<td>3.647</td>
<td>1.941</td>
<td>2.475</td>
</tr>
</tbody>
</table>
### Table 5: Prediction error on summary distribution statistics (demographic and human capital covariates model)

<table>
<thead>
<tr>
<th></th>
<th>Native workers</th>
<th></th>
<th>Immigrant workers</th>
<th></th>
<th>Cross-border workers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs.</td>
<td>Pred. error (%)</td>
<td>Obs.</td>
<td>Pred. error (%)</td>
<td>Obs.</td>
<td>Pred. error (%)</td>
</tr>
<tr>
<td></td>
<td>QR</td>
<td>S-M</td>
<td>Dag</td>
<td>LN</td>
<td>QR</td>
<td>S-M</td>
</tr>
<tr>
<td>Mean</td>
<td>23.99</td>
<td>-1.4</td>
<td>-0.9</td>
<td>-0.2</td>
<td>-1.2</td>
<td>18.56</td>
</tr>
<tr>
<td>Perc 10</td>
<td>11.55</td>
<td>1.7</td>
<td>5.3</td>
<td>6.7</td>
<td>6.0</td>
<td>9.30</td>
</tr>
<tr>
<td>Perc 20</td>
<td>14.43</td>
<td>-0.2</td>
<td>3.1</td>
<td>3.0</td>
<td>1.2</td>
<td>10.67</td>
</tr>
<tr>
<td>Perc 50</td>
<td>21.22</td>
<td>1.3</td>
<td>0.5</td>
<td>-0.5</td>
<td>-1.4</td>
<td>14.32</td>
</tr>
<tr>
<td>Perc 80</td>
<td>31.00</td>
<td>0.4</td>
<td>-1.5</td>
<td>-1.4</td>
<td>-0.6</td>
<td>23.81</td>
</tr>
<tr>
<td>Perc 90</td>
<td>37.56</td>
<td>-0.4</td>
<td>-1.0</td>
<td>0.7</td>
<td>2.0</td>
<td>32.26</td>
</tr>
<tr>
<td>SD</td>
<td>14.46</td>
<td>-17.3</td>
<td>-9.7</td>
<td>-7.9</td>
<td>-19.3</td>
<td>12.84</td>
</tr>
<tr>
<td>Gini</td>
<td>0.27</td>
<td>-5.4</td>
<td>-4.4</td>
<td>-2.5</td>
<td>-4.3</td>
<td>0.30</td>
</tr>
<tr>
<td>P90/P10</td>
<td>3.25</td>
<td>-2.1</td>
<td>-6.4</td>
<td>-6.0</td>
<td>-4.0</td>
<td>3.47</td>
</tr>
<tr>
<td>P90/P50</td>
<td>1.77</td>
<td>-1.7</td>
<td>-1.5</td>
<td>1.2</td>
<td>3.4</td>
<td>2.25</td>
</tr>
<tr>
<td>P10/P50</td>
<td>0.54</td>
<td>0.4</td>
<td>4.9</td>
<td>7.2</td>
<td>7.4</td>
<td>0.65</td>
</tr>
<tr>
<td>GE-1</td>
<td>0.12</td>
<td>-9.2</td>
<td>-3.9</td>
<td>-3.9</td>
<td>-9.7</td>
<td>0.14</td>
</tr>
<tr>
<td>GE0</td>
<td>0.12</td>
<td>-12.0</td>
<td>-7.6</td>
<td>-5.4</td>
<td>-11.6</td>
<td>0.14</td>
</tr>
<tr>
<td>GE1</td>
<td>0.13</td>
<td>-18.4</td>
<td>-11.6</td>
<td>-8.3</td>
<td>-18.2</td>
<td>0.17</td>
</tr>
<tr>
<td>GE2</td>
<td>0.18</td>
<td>-31.8</td>
<td>-17.6</td>
<td>-15.5</td>
<td>-36.1</td>
<td>0.24</td>
</tr>
</tbody>
</table>
Table 6: Prediction error on summary distribution statistics (demog., human capital and job char. covariates model)

<table>
<thead>
<tr>
<th></th>
<th>Native workers</th>
<th></th>
<th>Immigrant workers</th>
<th></th>
<th>Cross-border workers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs.</td>
<td>Pred. error (%)</td>
<td>Obs.</td>
<td>Pred. error (%)</td>
<td>Obs.</td>
<td>Pred. error (%)</td>
</tr>
<tr>
<td></td>
<td>QR</td>
<td>S-M</td>
<td>Dag</td>
<td>LN</td>
<td>QR</td>
<td>S-M</td>
</tr>
<tr>
<td>Mean</td>
<td>23.99</td>
<td>-1.0</td>
<td>-0.4</td>
<td>-0.1</td>
<td>-1.2</td>
<td>18.56</td>
</tr>
<tr>
<td>Perc 10</td>
<td>11.55</td>
<td>3.4</td>
<td>4.4</td>
<td>5.9</td>
<td>4.9</td>
<td>9.30</td>
</tr>
<tr>
<td>Perc 20</td>
<td>14.43</td>
<td>3.3</td>
<td>2.8</td>
<td>2.6</td>
<td>1.8</td>
<td>10.67</td>
</tr>
<tr>
<td>Perc 50</td>
<td>21.22</td>
<td>2.4</td>
<td>0.9</td>
<td>-0.2</td>
<td>-0.1</td>
<td>14.32</td>
</tr>
<tr>
<td>Perc 80</td>
<td>31.00</td>
<td>-1.5</td>
<td>-2.4</td>
<td>-2.2</td>
<td>-1.5</td>
<td>23.81</td>
</tr>
<tr>
<td>Perc 90</td>
<td>37.56</td>
<td>-2.0</td>
<td>-1.2</td>
<td>0.1</td>
<td>0.2</td>
<td>32.26</td>
</tr>
<tr>
<td>SD</td>
<td>14.46</td>
<td>-17.4</td>
<td>-4.4</td>
<td>-4.5</td>
<td>-19.2</td>
<td>12.84</td>
</tr>
<tr>
<td>Gini</td>
<td>0.27</td>
<td>-5.8</td>
<td>-2.9</td>
<td>-1.8</td>
<td>-4.8</td>
<td>0.30</td>
</tr>
<tr>
<td>P90/P10</td>
<td>3.25</td>
<td>-5.4</td>
<td>-5.7</td>
<td>-5.7</td>
<td>-4.7</td>
<td>3.47</td>
</tr>
<tr>
<td>P90/P50</td>
<td>1.77</td>
<td>-4.3</td>
<td>-2.1</td>
<td>0.3</td>
<td>0.3</td>
<td>2.25</td>
</tr>
<tr>
<td>P10/P50</td>
<td>0.54</td>
<td>1.1</td>
<td>3.5</td>
<td>6.1</td>
<td>5.0</td>
<td>0.65</td>
</tr>
<tr>
<td>GE-1</td>
<td>0.12</td>
<td>-1.5</td>
<td>-0.6</td>
<td>-1.8</td>
<td>-8.0</td>
<td>0.14</td>
</tr>
<tr>
<td>GE0</td>
<td>0.12</td>
<td>-9.0</td>
<td>-3.7</td>
<td>-3.2</td>
<td>-11.2</td>
<td>0.14</td>
</tr>
<tr>
<td>GE1</td>
<td>0.13</td>
<td>-17.5</td>
<td>-5.9</td>
<td>-4.8</td>
<td>-18.3</td>
<td>0.17</td>
</tr>
<tr>
<td>GE2</td>
<td>0.18</td>
<td>-32.8</td>
<td>-8.0</td>
<td>-8.8</td>
<td>-36.0</td>
<td>0.24</td>
</tr>
</tbody>
</table>
it at a fraction of estimation time. Furthermore, parametric models cut conditional quantile estimation time by between 91.5 and 98.5 percent in our application. These relative timings are indicative and will of course vary with alternative software, data and model specifications but we expect the timing gains of the maximum likelihood to remain considerable in other settings, to an extend that remains practically significant in full scale empirical applications involving for example resampling-based inference and detailed specification search. One size does not fit all however: no single best model emerges for the three sub-samples and the two sets of covariates.

6 Quantile wage gap decomposition

We now finally use our model parameters to estimate the components of the decomposition outlined in Eqs. (6) and (10). All estimates reported here are based on simulations from the Dagum parametric distribution specifications. The decompositions set out in Eqs. (6) and (10) are contingent on the choice of a baseline covariate distribution on which different ‘prices’ (conditional quantile predictions) are applied. The decompositions involve fixing a baseline covariate distribution and calculating what the wage distribution would look like if alternative conditional wage distributions were applied to these covariates. We took it to be the natives covariate distribution in the exposition in Section 2 and we start with this choice here thereby showing the magnitude of the wage advantage of the native workers. For the sake of completeness, we also conduct and report supra simulations applied to covariate distributions of immigrants as baseline (results with cross-border workers as baseline are available in Appendix C).

---

Timings for the estimation of the various models in our application were as follows. Since absolute timings reflect coding, the capacity of our software (4-core Stata/MP 12.1 for Windows (64-bit x86-64) updated to Feb 6 2012, running under Windows Server Enterprise) and hardware environments (Intel chipset Xeon CPU-E7 4850 @2GHZ 10-core with 128Gb of RAM), we normalize to 1 the timing for estimation of the quantile regression process based on a mesh of 99 equi-spaced quantiles on each of the three subgroups and for each of the two sets of covariates. Quantile regressions were estimated using repeated calls to Stata’s built-in qreg command. Prediction of conditional quantiles based on maximum likelihood estimation of the log-normal models took between 0.015 and 0.046 units of quantile regression (depending on subgroup and covariate set), where the absolute timings for quantile regression were between 115 and 644 seconds. The Dagum models took between 0.024 and 0.086 units of quantile regression and the Singh-Maddala model estimation took between 0.015 and 0.072 such units.

Appendix D reports results from alternative models. Despite the good fit to the empirical data of all models, small differences in model specifications have implications on counterfactual quantile estimates and therefore on resulting decomposition components, especially for decomposition of top quantiles. While the main substantive conclusions remain unaffected this observation calls for caution in interpreting decomposition results for extreme quantiles, even with large samples and when the models have satisfactory fit to the empirical distributions.
6.1 Native workers baseline

To start with, we present a set of results in which we account only for differences in demographic and human capital characteristics across the three groups: age, gender and educational achievement. Remember that cross-border workers are more predominantly men, younger, and with higher educational achievements, while immigrants are similar to natives but for their polarized educational records (many more immigrants have (at best) primary education or a tertiary degree).

Figure 4 shows (i), on top, the simulated quantile function among native workers (solid lines) and the counterfactual distributions that would be observed if these native workers were rewarded according to the parameters of the conditional quantile functions of immigrants (dashed lines) or of cross-border workers (dot-dashed lines); and (ii), at the bottom, the log difference between these two curves (which reflect the ‘unexplained’, ‘price’ component of the gap attributed to differences in the coefficients of the conditional quantile functions). The left side of the panel shows simulations in which the full foreign conditional quantile functions are applied upon native workers and therefore show the full ‘price’ effect. The right side of the panel shows the simulations that only adjust native workers’ wages to foreign workers mean wages and therefore shows the ‘levels’ part of the full ‘price’ effect. Note that we only report estimates between the 10th and the 90th percentile: more extreme percentile predictions become increasingly sensitive to model specification choices and display relatively large confidence intervals (all estimates are available on request).

The relative difference between the curves testifies of the non-linear difference between the wage distributions. The native worker advantage follows an inverted U shape. The unexplained gap between natives and foreign workers is low or negative for extreme top and bottom quantiles and peaks around the median for immigrant workers and around the 70th percentile for cross-border workers. Adjustment for human capital and demographic characteristics hardly reduces the magnitude of the distribution differences: observed quantiles of native workers wages generally remain between 0.10 and 0.30 log points larger than what they would be with foreign workers pay. The counterfactual quantile wage gaps remain close to the observed quantile wage gap so that the ‘composition’ effect is small.

A large part of the gap is accounted by differences in levels of pay. However, accounting only for differences in mean pay overestimates the wage advantage of native workers at both low and high quantiles. Comparison of Figures 4c and 4d reveals that foreign workers’ (conditional) wage distributions also differ in shape from native workers’ distributions: both low and high (conditional) quantiles among foreign workers are higher compared to the (conditional) means than similar quantiles among native workers. This is likely due at the bottom to the floor set at the minimum wage (which affects
native workers much less than foreign workers).\textsuperscript{11} At the top, the higher wage is a reflection of the ‘elite migration’ of high wage workers (Amétépé and Hartmann-Hirsch, 2011). This observation echoes the raw percentile ratios reported in Table 1. These relatively complex distributional differences across the three groups of workers lead to an advantage of native workers that is particularly marked for the middle 50 percent of native workers but that is much smaller (albeit still significantly positive) for the lowest and highest paid natives.

As we now show, sorting over different jobs and occupations accounts for a much bigger share of the wage gap than demographic and human capital differences. We include various job and occupation variables in the models to assess the wage gap with this richer set of controls. This accounts for the fact that native workers may be sorting themselves into better paying jobs, conditional on human capital characteristics, e.g., thanks to skills unobserved in our data such as language skills (Klein, 2003) or better job matches generally (Brosius et al., 2014, Brosius and Van Kerm, 2014). Native workers have much longer job tenure than foreign workers and generally work for large firms (1000+ employees). The sectoral and occupational patterns also vary across the three groups of workers, e.g., with immigrants over-represented in the construction sector or in low-skilled occupations, or natives much more likely to be employed as clerks or associate professionals and in the health, education or not-for-profit sector.

Accounting for the different job characteristics of foreign and native workers leads to a substantial decrease—although not an elimination—in the difference between counterfactual and observed wage distributions; see Figure 5. The pay advantage of native workers remains inverted-U shaped but at much lower levels than in Figure 4. It peaks at about 10 percent for the median native worker but it disappears compared to immigrant workers above the 70th percentile (above the 90th percentile compared to cross-border workers). The gap in fact turns significant and negative in favour of immigrant workers at the very top. The magnitude and significance of this effect at the very top however depends on the model specification (see Appendix D). Note that the ‘dispersion’ effect now becomes negligible as counterfactuals based on adjustment of means only are very similar to those based on full conditional distribution adjustments.

\footnote{In 2006, the minimum wage was at approximately 1500 euros per month, or just above 10 euros per hour on the basis of a standard full-time 144 hours per month employment.}
Figure 4: Native workers actual and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital and demographic controls only

Simulated quantile functions:

(a) Full conditional quantiles:
   full ‘price’ effect

(b) Rescaled conditional quantiles:
   ‘levels’ effect

Log difference between simulated quantile functions:

(c) Full conditional quantiles:
   full ‘price’ effect

(d) Rescaled conditional quantiles:
   ‘levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.
Figure 5: Native workers actual and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital, demographic and job characteristics controls

Simulated quantile functions:

(a) Full conditional quantiles

(b) Rescaled conditional quantiles: 'levels' effect

Log difference between simulated quantile functions:

(c) Full conditional quantiles: full ‘price’ effect

(d) Rescaled conditional quantiles: 'levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.
Table 7 reports log percentile wage gap decomposition components at selected percentiles as per Eqs (6) and (10). Results are also reported on the wage gap at the mean for comparison. Decomposition results confirm that a large part of the gap is accounted by compositional differences in job characteristics and by aggregate differences in wage levels: the largest components is the ‘levels’ when no occupations and job characteristics are included but becomes the ‘composition’ factor with the largest set of covariates. However the size and relative contribution of the different components vary with the percentile considered (dispersion effects are more important at the tails).

Interestingly, none of these results fully mimic the results obtained for the mean. The total gap for mean wage is generally lower than at the selected percentiles and, most notably, the unexplained ‘price’ effect after accounting for job characteristics is negligible at the mean (although it remains large and significant for much of the percentiles—especially the median—as Figure 5 illustrates). This testifies of the relevance of examining full distributional differences.

6.2 Foreign workers as baseline

For the sake of completeness, Figures 6 and 7 show figures similar to 4 and 5 when the baseline covariate distributions are taken to be those of immigrants. Mechanically, the main inverted-U shaped curves from Figures 4 and 5 flips into U-shaped curves: the advantage of native workers pay structure now reads as a disadvantage for immigrant and cross-border workers. Again, the gap is largely reduced once job and occupation characteristics are accounted for.

These figures are useful to examine wage differences between immigrants and cross-border workers. While mean pay is almost identical in the two groups, cross-border workers have higher wages up to approximately the 70th percentile and immigrants have higher wage percentiles above this point (see Figure 1). Figure 6 reveals that human capital and demographic characteristic differences account for around half of this gap. Figure 7 shows that the gap is further halved once job characteristics are considered and wage differences almost disappear between cross-border and immigrant workers in the bottom two-thirds of the distribution. However differences persist at the top: cross-border workers would have higher wages if they were rewarded as immigrants given their human capital and job characteristics. This observation is unlikely to reflect discriminatory mechanism. More likely, this reflects the endogenous mobility decisions of the highest paid workers from neighbouring countries to migrate.

---

12 Calculations for the mean are identical to those for quantiles barring the change in summary measure used. Mean and counterfactual means are calculated by taking the means over the pseudo-sample of stacked conditional quantile predictions.

13 Estimates with cross-border workers as baseline are provided in Figures C.3 and C.4 in Appendix C. They lead to qualitatively similar results.
### Table 7: Decomposition results at P10, P20, P50, P80, P90 and mean wage (native workers covariates as baseline)

<table>
<thead>
<tr>
<th></th>
<th>Simulated, HC only</th>
<th>Simulated, HC + Job char.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total (0) Compo. (1a) Price (2a)</td>
<td>Total (1b) Compo. (2b) Price (4b)</td>
</tr>
<tr>
<td></td>
<td>Levels (3a) Disp. (4a)</td>
<td>Levels (3b) Disp. (4b)</td>
</tr>
<tr>
<td>Natives vs. Immigrants log wage gap:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.256 (0.030)</td>
<td>0.261 (0.016) 0.239 (0.031) 0.000 (0.001)</td>
</tr>
<tr>
<td>Perc 20</td>
<td>0.302 (0.036)</td>
<td>0.309 (0.007) 0.273 (0.033) 0.016 (0.025)</td>
</tr>
<tr>
<td>Perc 50</td>
<td>0.394 (0.040)</td>
<td>0.387 (0.012) 0.274 (0.035) 0.078 (0.020)</td>
</tr>
<tr>
<td>Perc 80</td>
<td>0.264 (0.040)</td>
<td>0.308 (0.020) 0.253 (0.035) 0.031 (0.011)</td>
</tr>
<tr>
<td>Natives vs. Cross-border workers log wage gap:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.268 (0.022)</td>
<td>0.281 (0.011) 0.279 (0.024) 0.001 (0.000)</td>
</tr>
<tr>
<td>Perc 20</td>
<td>0.213 (0.034)</td>
<td>0.231 (0.005) 0.252 (0.026) (-0.009 (0.022) )</td>
</tr>
<tr>
<td>Perc 50</td>
<td>0.327 (0.037)</td>
<td>0.321 (0.007) 0.272 (0.027) 0.057 (0.020)</td>
</tr>
<tr>
<td>Perc 80</td>
<td>0.291 (0.029)</td>
<td>0.315 (0.014) 0.294 (0.027) 0.022 (0.010)</td>
</tr>
</tbody>
</table>

Notes: Column (0) reports the gap observed in the raw data. Columns (1a) and (1b) report the gap as per model-based simulations and the decompositions are (1a)=(2a)+(3a)+(4a) and (1b)=(2b)+(3b)+(4b); see Section 2 for details on the decomposition. The first set refers to a model with human capital and demographic characteristics only. The second set refers to a model including job characteristics too. Bootstrap standard errors in brackets.
into Luxembourg rather than remain cross-border workers (they may be more able to afford migration cost related, e.g., to housing and may have stronger local labour market attachment too).

Decomposition results are reported in Table 8 for the immigrant vs. cross-border workers wage gap. Again, results illustrate how differentials in mean wage may hide contrasted variations at percentiles.

Table 8: Decomposition results at P10, P20, P50, P80, P90 and mean wage (immigrant workers covariates as baseline)

<table>
<thead>
<tr>
<th>Obs.</th>
<th>Simulated, HC only</th>
<th>Simulated, HC + Job char.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Compo.</td>
</tr>
<tr>
<td></td>
<td>Levels</td>
<td>Disp.</td>
</tr>
<tr>
<td>Immigrants vs. Cross-border workers log wage gap:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.011 (0.021)</td>
<td>0.020 (0.010)</td>
</tr>
<tr>
<td>Perc 20</td>
<td>-0.089 (0.013)</td>
<td>-0.078 (0.007)</td>
</tr>
<tr>
<td>Perc 50</td>
<td>-0.066 (0.017)</td>
<td>-0.066 (0.011)</td>
</tr>
<tr>
<td>Perc 80</td>
<td>0.027 (0.027)</td>
<td>0.008 (0.012)</td>
</tr>
</tbody>
</table>

Notes: Column (0) reports the gap observed in the raw data. Columns (1a) and (1b) report the gap as per model-based simulations and the decompositions are (1a)=(2a)+(3a)+(4a) and (1b)=(2b)+(3b)+(4b); see Section 2 for details on the decomposition. The first set refers to a model with human capital and demographic characteristics only. The second set refers to a model including job characteristics too. Bootstrap standard errors in brackets.
Figure 6: Immigrant workers actual and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital and demographic controls only

Simulated quantile functions:

(a) Full conditional quantiles: full ‘price’ effect

(b) Rescaled conditional quantiles: ‘levels’ effect

Log difference between simulated quantile functions:

(c) Full conditional quantiles: full ‘price’ effect

(d) Rescaled conditional quantiles: ‘levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.
Figure 7: Immigrant workers actual and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital, demographic and job characteristics controls

Simulated quantile functions:

(a) Full conditional quantiles:
   full ‘price’ effect

(b) Rescaled conditional quantiles:
   ‘levels’ effect

Log difference between simulated quantile functions:

(c) Full conditional quantiles:
   full ‘price’ effect

(d) Rescaled conditional quantiles:
   ‘levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.
7 Summary and conclusion

This paper set out to examine private sector wage differences between native, immigrant and cross-border workers in Luxembourg. Along the way, we describe and assess the performance of a computationally convenient alternative estimator of the Machado and Mata (2005) decomposition of quantile wage gaps into ‘price’ and ‘composition’ components.

The decomposition of quantile wage gaps has become popular in analyses of wage differentials (see Fortin et al., 2011). Our application demonstrates that the Machado and Mata (2005) algorithm—and its variants described in Autor et al. (2005) or Melly (2005)—can be closely approximated at a fraction of the computational cost by substituting estimation of a large number of quantile regressions by a single “parametric quantile regression”. This involves specifying a conventional parametric wage distribution model—in our application the log-normal, Dagum or Singh-Maddala distributions—with parameters allowed to vary flexibly with covariates (Biewen and Jenkins, 2005, Van Kerm, 2013a) and deriving all necessary quantile predictions from a small set of model parameters. Surprisingly, the fit of the parametric models to our empirical distributions is generally better than the fit achieved with a sequence of 99 quantile regressions, in particular with a large set of covariates. This alternative estimator should therefore appeal to practitioners since its much lower computational cost makes resampling-based inference tractable and facilitates practical robustness and specification checks. The validity of parametric specifications also opens possibilities to address endogenous selection using familiar likelihood-based models (Pigini, 2012, Christofides et al., 2013, Van Kerm, 2013a).

While all models provide satisfactory fit to the data, we find no clear ordering among the three parametric models considered; their relative performance varies with sub-sample, covariates and according to the fit statistics examined. This last observation commends the use of multiple fit indicators and reminds of the difficulty in identifying a single ‘optimal’ model fit for all quantile points of interest. Model averaging may be an avenue of research for combining estimates returned from multiple models (Claeskens and Hjort, 2008).

Substantive results point to a clear inverted U shape for the ‘native worker advantage’ in the private sector: the advantage is low and possibly negative for both low and high quantiles, but it is large, positive and significant for the middle half of the quantile range (between the 20th and the 70th percentiles approximately). The pattern holds against both immigrants and cross-border workers and persists after controlling for differences in human capital and job characteristics. Our data confirm that, e.g.,

---

14Evidence as to whether the gap effectively becomes significantly negative above the 90th percentile of the native workers distribution is one result that varies across model parameterizations.
cross-border workers have generally higher educational achievements while immigrants are polarized at both ends of the education distribution, but the estimated effect of these compositional differences on the wage distribution differences is small. We note however that controlling for differences in job characteristics largely reduces—but does not eliminate—the native worker advantage. The absence of a wage gap at low quantiles is in contrast with results observed recently in Spain (Simón et al., 2008) and the UK (Hunt, 2012) and can be ascribed to the strict minimum wage legislation that compress potential differences at low wages. The evidence of a catching up or even a negative gap (an ‘immigrant worker advantage’) at the top is consistent with evidence of the high-skill, ‘elite’ migration described in Amétépé and Hartmann-Hirsch (2011) but is also featured in other countries.

Inspection of the whole quantile functions turns out to be particularly informative of differences in pay between cross-border and immigrant workers. While average pay is very similar across the two groups, the quantile function is much flatter among cross-border workers. In particular it does not exhibit as much catching up to top native wage levels. After controlling for differences in demographic, human capital and job characteristics, pay turns out to be almost equal for much of the distribution. All difference that remains between the two groups of foreign workers is at the top. This is likely to partly reflect endogenous settlement decisions of foreign workers from neighbouring countries, with highest paid workers more likely to accept the cost of migration in Luxembourg.

Needless to say, the ‘unexplained’ pay advantage in favour of native workers need not reflect discrimination. It is possible to ascribe the different sorting of foreign workers across different occupations and the unexplained difference in wage levels to other factors such as unobserved productive characteristics (e.g., multiple language proficiency is relevant in the local context, Klein, 2003), workers’ preferences for different employers (Brosius et al., 2014), asymmetric wage bargaining power—especially at entry in Luxembourg (Brosius, 2005). The extent to which these factors account for the sizeable unexplained wage gap in the middle of the wage distribution remains however unknown.

References


StataCorp (2011). *Stata Statistical Software: Release 12*. StataCorp LP, College Station.


Appendix A  Technical annex: implementation details

This technical annex provides details on model specifications and practical implementation steps.

Model specifications

The full vector of human capital and demographic characteristics that we condition upon in our models includes a dummy for gender, a linear spline in age interacted with the gender dummy (with knots at ages 30, 40 and 50) and two education-level dummies. Job characteristics include a dummy for part-time employment, a dummy for jobs involving supervision of other workers, a linear spline in job tenure (with knots at 2, 5 and 10 years with the current employer), four firm size dummies, seven sector dummies, and seven occupation dummies (see Table 2 for their labels). A restricted vector does not include the gender \times age interaction and the supervisory role dummy.

Quantile regressions include the full vector of covariates in \( P(X) \) in (11). The log-normal models include the full vector for the \( m \) parameter and the restricted vector for the variance parameter \( s^2 \) in (13). The Singh-Maddala models include the full vector for the scale parameter \( b \), and the restricted vector for \( a \) and \( q \) in (14). Analogously, the Dagum models include the full vector for the scale parameter \( b \), and the restricted vector for \( a \) and \( p \) in (15). Restricting the vector set in the likelihood-based models improved convergence with little restriction on the specification.

Top and bottom coding of extreme predictions

To avoid a small number of outlying predictions to exert excessive influence on extreme unconditional quantile predictions, we top-coded all conditional quantile predictions obtained from equations (11), (13), (14) or (15) with our model parameters to 250 euros per hour and bottom-coded all predictions to 5 euros per hour. These values correspond to the minimum and maximum hourly wage values allowed in the data. These are wide bounds since 5 euros per hour is half the legal minimum wage and 250 euros is more than 10 times mean native workers wage. These safe-guarding adjustments practically affect only the mean and the most extreme quantile predictions.

Sub-sampling from simulated pseudo-samples

Calculation of unconditional (counterfactual) quantiles \( Q_{\langle j \mid k \rangle}(\tau) \) and \( \tilde{Q}_{\langle j \mid k \rangle}(\tau) \) by the Autor et al. (2005) and Melly (2005) procedure described in Section 2 involves calculating quantiles in a simulated pseudo-sample of \( n_k \times K \) observations. With our sample sizes and a mesh of 99 conditional quantile predictions, this corresponds to vectors of sizes 678, 843, 750, 519 and 1, 373, 625 when pre-
dicting on natives, immigrants or cross-border workers covariate baselines respectively. To speed up calculations, we estimated unconditional quantiles from random sub-samples of 25 percent of these vectors. To preserve the distribution of covariates and the validity of model comparisons, sub-sampling was conducted separately from each observation’s 99 conditional quantile predictions and the same conditional quantile predictions were selected from all models. This procedure does not bias unconditional quantile predictions but potentially inflates their variance. Sub-sampling was repeated in each bootstrap replication (see below) so the variability introduced in the procedure is captured in bootstrap confidence intervals and standard errors.

**Bootstrap inference**

Chernozhukov et al. (2013) recommend bootstrap inference and demonstrate the validity of exchangeable bootstrap procedures for inference on counterfactual distributions. We implemented the repeated half-sample bootstrap of Saigo et al. (2001), accounting for the two-stage sampling design of the data described in Section 4. Five-hundred Bootstrap replicates of all conditional and (counterfactual) unconditional quantiles estimates (and all components of Eq. (6)) were computed on the basis of bootstrap replication weights as described in Chernozhukov et al. (2013) and pointwise confidence intervals shown in Section 6 were constructed with the bias-corrected percentile method (Efron, 1981). We use the *rhsbsample* Stata user-written command for generating the replication weights (Van Kerm, 2013b).
Appendix B  Observed and simulated unconditional quantiles: prediction error
Figure B.1: Log difference between observed and predicted unconditional quantile functions (with 95 percent confidence bootstrap variability bands for parametric models)

Demographic and human capital covariates only:

(a) Native workers

(b) Immigrant workers

(c) Cross-border workers

Demographic, human capital and job characteristics:

(d) Native workers

(e) Immigrant workers

(f) Cross-border workers

Note: Vertical bars show 95 percent bias-corrected percentile-based bootstrap variability bands based on 500 repeated half-sample bootstrap replications.
Figure B.2: Absolute log difference between observed and predicted unconditional quantile functions (with 95 percent confidence bootstrap variability bands for parametric models)

Demographic and human capital covariates only:

(a) Native workers
(b) Immigrant workers
(c) Cross-border workers

Demographic, human capital and job characteristics:

(d) Native workers
(e) Immigrant workers
(f) Cross-border workers

Note: Vertical bars show 95 percent bias-corrected percentile-based bootstrap variability bands based on 500 repeated half-sample bootstrap replications.
Appendix C  Decomposition results with cross-border workers as baseline
Figure C.3: Cross-border workers actual and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital and demographic controls only

Simulated quantile functions:

(a) Full conditional quantiles: full ‘price’ effect

(b) Rescaled conditional quantiles: ‘levels’ effect

Log difference between simulated quantile functions:

(c) Full conditional quantiles: full ‘price’ effect

(d) Rescaled conditional quantiles: ‘levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.
**Figure C.4:** Cross-border workers actual and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital, demographic and job characteristics controls

**Simulated quantile functions:**

(a) Full conditional quantiles: full ‘price’ effect

(b) Rescaled conditional quantiles: ‘levels’ effect

---

**Log difference between simulated quantile functions:**

(c) Full conditional quantiles: full ‘price’ effect

(d) Rescaled conditional quantiles: ‘levels’ effect

---

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.
Appendix D  Counterfactual quantiles and quantile wage gap decomposition with alternative model specifications (for native workers baseline)
Figure D.5: Native workers observed and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital and demographic controls only (quantile regression)

**Simulated quantile functions:**

(a) Full conditional quantiles: 
  full ‘price’ effect

(b) Rescaled conditional quantiles: 
  ‘levels’ effect

Log difference between the simulated quantile functions:

(c) Full conditional quantiles: 
  full ‘price’ effect

(d) Rescaled conditional quantiles: 
  ‘levels’ effect

Note: Light grey background lines in bottom panel show total observed gap.
Figure D.6: Native workers observed and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital and demographic controls only (Dagum distribution)

Simulated quantile functions:

(a) Full conditional quantiles: full ‘price’ effect

(b) Rescaled conditional quantiles: ‘levels’ effect

Log difference between the simulated quantile functions:

(c) Full conditional quantiles: full ‘price’ effect

(d) Rescaled conditional quantiles: ‘levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). Light grey background lines in bottom panel show total observed gap.
Figure D.7: Native workers observed and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital and demographic controls only (Singh-Maddala distribution)

Simulated quantile functions:

(a) Full conditional quantiles: full ‘price’ effect

(b) Rescaled conditional quantiles: ‘levels’ effect

Log difference between the simulated quantile functions:

(c) Full conditional quantiles: full ‘price’ effect

(d) Rescaled conditional quantiles: ‘levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.
Figure D.8: Native workers observed and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital and demographic controls only (log-normal distribution)

Simulated quantile functions:

(a) Full conditional quantiles: full ‘price’ effect

(b) Rescaled conditional quantiles: ‘levels’ effect

Log difference between the simulated quantile functions:

(c) Full conditional quantiles: full ‘price’ effect

(d) Rescaled conditional quantiles: ‘levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.
Figure D.9: Native workers observed and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital, demographic and job characteristics controls (quantile regression)

Simulated quantile functions:

(a) Full conditional quantiles: full ‘price’ effect

(b) Rescaled conditional quantiles: ‘levels’ effect

Log difference between the simulated quantile functions:

(c) Full conditional quantiles: full ‘price’ effect

(d) Rescaled conditional quantiles: ‘levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.
Figure D.10: Native workers observed and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital, demographic and job characteristics controls (Dagum distribution)

Simulated quantile functions:

(a) Full conditional quantiles: full ‘price’ effect

(b) Rescaled conditional quantiles: ‘levels’ effect

Log difference between the simulated quantile functions:

(c) Full conditional quantiles: full ‘price’ effect

(d) Rescaled conditional quantiles: ‘levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.
Figure D.11: Native workers observed and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital, demographic and job characteristics controls (Singh-Maddala distribution)

Simulated quantile functions:

(a) Full conditional quantiles: full ‘price’ effect

(b) Rescaled conditional quantiles: ‘levels’ effect

Log difference between the simulated quantile functions:

(c) Full conditional quantiles: full ‘price’ effect

(d) Rescaled conditional quantiles: ‘levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.
Figure D.12: Native workers observed and counterfactual wage quantiles (top) and log wage gap estimates with confidence intervals (bottom): human capital, demographic and job characteristics controls (log-normal distribution)

Simulated quantile functions:

(a) Full conditional quantiles: full ‘price’ effect

(b) Rescaled conditional quantiles: ‘levels’ effect

Log difference between the simulated quantile functions:

(c) Full conditional quantiles: full ‘price’ effect

(d) Rescaled conditional quantiles: ‘levels’ effect

Note: Vertical spikes show 95 percent bias-corrected percentile-based bootstrap variability bands (based on 500 repeated half-sample bootstrap replications). For readability, variability bands are shown at every four percentiles only. Light grey background lines in bottom panel show total observed gap.