This is a pre-copyedited, author-produced version of an article accepted for publication in Journal of the Royal Statistical Society Series A: Statistics in Society, following peer review. The version of record [M. J. Lombardía, E. López-Vizcaíno, y C. Rueda, «A New Approach to the Gender Pay Gap Decomposition by Economic Activity», Journal of the Royal Statistical Society Series A: Statistics in Society, vol. 185, n.º 1, pp. 219-245, ene. 2022] is available online at: https://doi.org/10.1111/rssa.12742.

A new approach to the gender pay gap decomposition by economic activity*

María José Lombardía[1], Esther López-Vizcaíno[2]

Cristina Rueda[3]

[1] Universidade da Coruña, CITIC, Spain,
[2] Instituto Galego de Estatística, Spain,
[3]Universidad de Valladolid, Spain

Abstract

The aim of this paper is to present an original approach to estimate the gender pay gap. We propose a model-based decomposition, similar to the most popular approaches, where the first component measures differences in group characteristics and the second component measures the unexplained effect; the latter being the real gap. The novel approach incorporates model selection and bias correction. The pay gap problem in a small area context is considered in this paper, although the approach is flexible to be applied to other contexts.

Specifically, the methodology is validated for analysing wage differentials by economic activities in the region of Galicia (Spain) and by analysing simulated data from an experimental design that imitates the generation of real data. The good performance of the proposed estimators is shown in both cases, specifically when compared with those obtained from the widely used Oaxaca-Blinder approach.

1 Introduction

Gender wage inequality is a documented fact that happens in almost all of the industrialized countries (Juhn et al., 1993, Oaxaca and Ransom, 1994,1998, Aláez and Ullibarri, 2001, Moral-Arce et al., 2011, Fortin et al., 2011) and interest in the quantification of this gap has increased in recent years.

In the case of the Spanish labor market, a number of actions were taken to provide women with real access to employment with full social and economic rights, with special emphasis on the reconciliation of family and working life. This and other reasons led to a massive incorporation of women into the labor market, but it has not resulted in equality, that is, women's unemployment rates are higher in most sectors, the jobs that women take do not involve the same degree of responsibility or decision-making power and women's participation is limited to a few sectors of the economy. But more important is that, for a similar job, it has been shown that men have better wages than women (Fernández et al., 2000; Aláez and Ullibarri, 2001;

*Supported by the MINECO grants MTM2017-82724-R, MTM2015-71217-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C 2020/14 and Centro de Investigación del Sistema universitario de Galicia ED431G 2019/01), all of them through the ERDF.

Aláez et al., 2000-2003; De la Rica et al., 2008; Moral-Arce et al., 2011).

The gender pay gap (GPG), calculated as the difference between men's average hourly earnings and women's average hourly earnings, usually expressed as a percentage of the men's average hourly earnings, is widely used as the key indicator to study the progress in the European Union (EU), see Leythienne and Ronkowski (2018), among others. Moreover, most researchers in this field (Jann, 2008, Moral-Arce et al., 2011, Anastasiade and Tillé, 2017a, 2017b, Hlavac, 2018) work with the transformed variable to logarithms, $Y = log(W)$, where W is the hourly wages. The rationale behind the use of the logarithmic transformation is that the salaries have a skew distribution. However, the skewness is also exhibited on the log scale. Finding the optimal transformation is not straightforward. As an alternative, Graf et al. (2019) propose considering a much more flexible distribution called generalized beta of the second kind (GB2). In this work, we follow the ideas of most researchers. We are aware that it would be interesting to incorporate this proposal, but it goes beyond the objectives of this paper.

Typically the sample estimators are used directly as the GPG measure because the sample size are often very high. Instead, a more formal definition of the GPG is used here, as we deal with scenarios with small sample sizes. It is given, using expectations, as follows:

$$GPG = \frac{E(W_m) - E(W_w)}{E(W_m)},$$

where the subscripts m are the men's wage and w the women's wage.

The most widely used approach to analyze the gender gap is the decomposition due to Oaxaca (1973) and Blinder (1973); Specifically, Eurostat applies this approach to decompose the estimator of GPG (Leythienne and Ronkowski, 2018). The Oaxaca-Blinder (OB) decomposition breaks down the difference $\Delta = E(Y_m) - E(Y_w)$ into two components: $\Delta = Q + U$, where the first component explains the difference between the observed productive characteristics, such as education or work experience, and the second accounts for differences in the structure of the model (Anastasiade and Tillé, 2017a, 2017b). The U component is known as the unexplained component and is usually considered to be the wage discrimination by gender in the labor market. By instances, Popli (2013) states that the unexplained wage gap includes the effect of labour market discrimination, unobservable variables and omitted variables. The methodology presented in this paper accounts for the bias from unobservable or omitted variables.

So, the GPG decomposition is:

$$GPG_Q = GPG\frac{Q}{\Delta} \quad \text{and} \quad GPG_U = GPG\frac{U}{\Delta}, \tag{1}$$

which represent the explained and unexplained part of the GPG, respectively. The OB method, including the definitions of Q and U, is briefly revised in Section 2.

Despite its widespread use, the standard Oaxaca-Blinder decomposition has important limitations; firstly, too simple parametric functions are often used for the regression model; secondly, the results may depend heavily on the set of explanatory variables selected; and thirdly, the unexplained variability is not always taken into account and generates bias.

This paper aims to adapt the methodology to estimate GPG_Q and GPG_U in small areas, which are disaggregated levels of the population, and specifically analyze the wage differentials by economic activities in the region of Galicia (Spain). Often in small area contexts, the sample sizes are too small, which implies the standard estimators are unreliable. In particular, in our application for 25% of the economic activities, we have information on less than 26 male workers, being the sample sizes for women even smaller.

In fact, no official estimators of wages are provided at the economic activity level because direct estimates have very low accuracy. The latter is a typical SAE (Small Area Estimation) problem. The most extended SAE approach is to give a model-based estimator, where explanatory variables is incorporated. Besides, the model is often defined as a mixed effect model where a random effect represents the areas (see Rao and Molina(2015).

The proposal here is to combine SAE methodology and OB methodology by deriving a decomposition inspired by (1), adapted to a small area scenario in such a way that for a given area d, GPG_d is also decomposing into the GPG_{Qd} and GPG_{Ud}. These components are efficiently estimated using an approach in three steps that use SAE techniques and overcome OB's drawback, commented above. First, by doing model selection using an AIC criterion, and second, by including a bias correction term that accounts for omitted variable bias as in many applications, the implicit assumption that all confounding variables are included is not easy to attain. Confidence intervals for GPG_{Qd} and GPG_{Ud} are obtained for each area.

Furthermore, to validate the methodology proposal, a simulation study is conducted. It has a particular design that mimics the generation of data of the real application at hand. The estimators obtained with the novel approach are compared with those using the OB approach, showing that the formers outperform the latter.

We organize the remainder of the paper as follows. Section 2 revises the OB approach and SAE models, while Section 3 details the novel proposal. The estimation of the gender gap by economic activities in the region of Galicia (Spain) is presented in Section 4 and the simulation study in Section 5. Finally, the main conclusions and a brief discussion of future research is given in Section 6.

2 Background

We use the notation which is usually considered for the OB decomposition and SAE models. The beta parameter appears in both formulations, but with a different role in each one.

2.1 Pay gap decomposition

Let $Y_g = log(W_g)$ be the natural log of hourly earnings for $g = m, w$; men and women, respectively. And let $\mathbf{X}_g$;$(X_{g1}, \ldots, X_{gp})$ represent the set of available explanatory variables. Typically, these explanatory variables are the education level, the age of the worker or the work experience.

Assume that samples of sizes n_m, n_w are available for men and women, respectively, and the following models:

$$Y_{gi} = \mathbf{X}_{gi}\boldsymbol{\beta}_g + \epsilon_{gi}, \quad i = 1, ..., n_g, \quad g \in (m, w);$$

where $\boldsymbol{\beta}_g = (\beta_{g1}, \ldots, \beta_{gp})'$ is an unknown vector of parameters and ϵ_{gi} is independent, with $E(\epsilon_g) = 0$.

Then,

$$\Delta = E(Y_m) - E(Y_w) = \overline{\mathbf{X}}_m\boldsymbol{\beta}_m - \overline{\mathbf{X}}_w\boldsymbol{\beta}_w,$$

where, $\bar{\mathbf{X}}_m$ and $\bar{\mathbf{X}}_w$ are the corresponding sample means.

Now, let β_r be the auxiliary coefficient; then, adding and subtracting $\overline{\mathbf{X}}_m\boldsymbol{\beta}_r$ and $\overline{\mathbf{X}}_w\boldsymbol{\beta}_r$, the above difference can be written as:

$$\Delta = (\overline{\mathbf{X}}_m - \overline{\mathbf{X}}_w)\boldsymbol{\beta}_r + \overline{\mathbf{X}}_m(\boldsymbol{\beta}_m - \boldsymbol{\beta}_r) + \overline{\mathbf{X}}_w(\boldsymbol{\beta}_r - \boldsymbol{\beta}_w).$$

Frequently, the discrimination is directed toward one of the groups (women, in our case), which determines the value of the auxiliary coefficient, so that $\beta_r = \beta_m$ and then,

$$\Delta = (\overline{\mathbf{X}}_m - \overline{\mathbf{X}}_w)\boldsymbol{\beta}_m + \overline{\mathbf{X}}_w(\boldsymbol{\beta}_m - \boldsymbol{\beta}_w).$$

where

$$Q = (\overline{\mathbf{X}}_m - \overline{\mathbf{X}}_w)\boldsymbol{\beta}_m,$$

is the component that is explained by group differences in the predictors (the "quantify effect" or "the explained part") and

$$U = \overline{\mathbf{X}}_w(\boldsymbol{\beta}_m - \boldsymbol{\beta}_w).$$

is the unexplained component, usually attributed to discrimination. Notice that U also captures all the potential effects of differences in unobserved variables.

The estimation of the decomposition is straightforward. Let $\hat{\boldsymbol{\beta}}_m$ and $\hat{\boldsymbol{\beta}}_w$ be the least-squares estimates for $\boldsymbol{\beta}_m$ and $\boldsymbol{\beta}_w$ obtained separately from the two samples. Then,

$$\begin{aligned}
\hat{\Delta} &= \bar{Y}_m - \bar{Y}_w = \hat{Q} + \hat{U}, \\
\hat{Q} &= (\bar{\mathbf{X}}_m - \bar{\mathbf{X}}_w)\hat{\boldsymbol{\beta}}_m, \\
\hat{U} &= \bar{\mathbf{X}}_w'(\hat{\boldsymbol{\beta}}_m - \hat{\boldsymbol{\beta}}_w).
\end{aligned}$$

More details and other decompositions can be seen in Jann (2008) and Hlavac (2018).

2.2 SAE models

The wage estimation by small areas is a typical SAE problem as it is assumed that the sample sizes are too low in many areas. The most extended approach in SAE problems is to give a model-based estimator that usually includes random effects due to the areas. To clarify the presentation we have only considered linear models, as these are widely used in SAE applications and are well suited to analyzing the problem at hand. Nevertheless, the proposal in the paper can be adapted to consider other additive non-linear models as those which include monotone and spline models such as those in Opsomer et al. (2008) and Rueda and Lombardía (2012).

Specifically, the so-called nested error regression model with sampling weights, which, for the i-th worker in area d, is defined as follows:

$$Y_{di} = \mathbf{X}_{di}\boldsymbol{\beta}_d + \mathbf{Z}_{di}\mathbf{u}_d + \epsilon_{di}, d = 1, \ldots, D, i = 1, \ldots, n_d; \tag{2}$$

where Y_{di} is the (log) wage, $\mathbf{X}_{di} = (X_{1di}, \ldots, X_{pdi})$ is a p vector of explanatory variables, $\boldsymbol{\beta}_d$ is the vector of unknown parameters; $\mathbf{Z}_{di}\mathbf{u}_d$ are the random effects, with $\mathbf{Z}_{di}$ being a q dimensional vector also defined from the explanatory variables, and $\mathbf{u}_d \in N(0, \boldsymbol{\Sigma}_u)$, where $\boldsymbol{\Sigma}_u$ is a matrix of dimension q. $\mathbf{u}_d$ is independent of the model error ϵ_{di}, which are assumed to be independent

$N(0, \sigma_e^2 w_{di}^{-1})$, where w_{di} is the sampling weight for individual i in area d. In these models the covariance of the response variable in area d is $\mathbf{V}_d = \mathbf{Z}_d \mathbf{\Sigma}_u \mathbf{Z}_d + \mathbf{\Sigma}_e$ where $\mathbf{\Sigma}_e = \sigma_e^2 W_d{}^{-1}$ and $W_d = diag(w_{di})_{n_d \times n_d}$. Two particular interesting cases are considered by Battese et al. (1988) and Dempster et al. (1981).

In SAE applications, researchers are mainly interested in deriving estimators for the response variable in small areas; they usually work with the conditional means as the parameters of interest, which are defined as follows:

$$\mu_d = E(Y_d|\mathbf{u}_d) = \bar{\mathbf{X}}_d \boldsymbol{\beta}_d + \bar{\mathbf{Z}}_d \mathbf{u}_d, d = 1, \ldots, D$$

The estimation process starts with the variance components $(\mathbf{\Sigma}_u, \mathbf{\Sigma}_e)$, which are estimated by the residual maximum likelihood (REML) method; the empirical best linear unbiased estimators of $\boldsymbol{\beta}_d$ and the empirical best linear unbiased predictors of $\mathbf{u}_d$ are then obtained. Furthermore, estimators for μ_d are derived as follows:

$$\hat{\mu}_d = \bar{\mathbf{X}}_d \hat{\boldsymbol{\beta}}_d + \bar{\mathbf{Z}}_d \hat{\mathbf{u}}_d, d = 1, \ldots, D$$

For details on the estimation process and further learning about SAE methods, the reader is referred to the monographs of Jiang (2007) and Rao and Molina (2015).

3 The novel proposal

Assume that we have wage data for men and women and explanatory variables from samples from D small areas, and that we are interested in deriving estimates for the two components of the GPG.

A three-step approach is proposed to derive GPG estimators. The first step is a model selection step; in the second step, preliminary area-specific estimators for GPG_Q and GPG_U are obtained from a GPG decomposition adapted to small areas; and finally, in the third step, a Monte-carlo algorithm is consider to derive bias corrected GPG_Q and GPG_U estimators and confidence intervals.

First Step: Model Selection

There are different approaches to model selection, and the task is complicated when random effects are present. The discussions by Lombardia et al. (2017), where the GAIC is first presented, and by Fan and Li (2012), where iterative regularization methods are proposed, are very instructive. Besides proposing a method for making the model selection, a list of candidate models must be defined. From the methodological point of view, it can be as large as the practitioner decides. From the computational and interpretative points of view, a reduced list of candidate models are recommended.

The strategy adopted in this paper is to define a reduced list of candidate models and then to use the GAIC statistic to select a model from this list. In particular, for applications with a reduced number and well-known, explanatory variables, we recommend the use of *prior knowledge* to decide an initial selection of fixed and random effects. Typically, in small area applications, the variable determining the areas is modelled using random effects. Other candidate models may be defined by considering the interactions of that variable, with other explanatory variables.

The group (men/women) with the larger sample size is used to select the model. In *GPG* applications, it is usually that of men (m). The models from which we make the choice differ in the explanatory variables selected and the type of effects (fixed or random) that describe the model and are formulated as the general model (2), as follows:

$$M : Y_{mdi} = \mathbf{X}_{mdi}\boldsymbol{\beta}_{md} + \mathbf{Z}_{mdi}\mathbf{u}_{md} + \epsilon_{mdi}, \qquad d = 1, \dots, D, i = 1, \dots, n_d.$$

The selection of the best model is done using an *AIC* statistic (Akaike, 1973); specifically, we use an adapted version for unit level models of the *xGAIC* proposed by Lombardía et al. (2017) with the following expression for the model M

$$xGAIC = -2log(l_x(M)) + xGDF$$

where $l_x(M)$ is the quasi-loglikelihood of the model M, which considers the focus on the random effect and the total variability and $xGDF$ is a measure of the complexity of the model, as follows:

$$\log(l_x(M)) = -\frac{1}{2}D\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_{Y_m}| - \frac{1}{2}(\mathbf{Y}_m - \boldsymbol{\mu}_m)'\mathbf{V}_{Y_m}^{-1}(\mathbf{Y}_m - \boldsymbol{\mu}_m)$$

and

$$xGDF = \sum_{d=1}^{D}\frac{\partial E(\widehat{\boldsymbol{\mu}}_{md})}{\partial \boldsymbol{\mu}_{md}} = \sum_{d=1}^{D}\sum_{i=1}^{n_{md}}\sum_{j=1}^{n_{md}} V_{mdij}^{-1}Cov(\widehat{\mu}_{mdi}, Y_{mdj})$$

where $\mathbf{Y}_m$ is the $\sum_{d=1}^{D} n_d \times 1$ vector with elements Y_{mdi}, $\boldsymbol{\mu}_m$ is the vector of the conditional means $\boldsymbol{\mu}_m = E(\mathbf{Y}_m|\mathbf{u})$ with elements $\boldsymbol{\mu}_{mdi} = \mathbf{X}_{mdi}\boldsymbol{\beta}_{md} + \mathbf{Z}_{mdi}\mathbf{u}_{md}$, $\mathbf{V}_{Y_m}$ is the conditional covariance matrix, which is diagonal with elements $\mathbf{V}_{md}$ according to Section 2.2 and V_{mdij}^{-1} is the ij-element of $\mathbf{V}_{md}^{-1}$.

Parametric bootstrap is used to estimate $xGDF$, as the analytic values are difficult to obtain in a similar way to Lombardía et al. (2017). The model with the lowest value of $xGAIC$ is selected.

Second Step: GPG decomposition. Preliminary estimators.

Let X_m, Z_m and X_w,Z_w be the design matrices representing the explanatory variables and the random effects on the selected model for men and women, respectively. For a given area, d, consider $\Delta_d = \mu_{md} - \mu_{wd}$, which is decomposed as follows:

$$\Delta_d = \mu_{md} - \mu_{wd} = \bar{\mathbf{X}}_{md}\boldsymbol{\beta}_{md} + \bar{\mathbf{Z}}_{md}\mathbf{u}_{md} - \bar{\mathbf{X}}_{wd}\boldsymbol{\beta}_{wd} - \bar{\mathbf{Z}}_{wd}\mathbf{u}_{wd}.$$

$$\begin{aligned}
\Delta_d &= Q_d + U_d \\
Q_d &= (\bar{\mathbf{X}}_{md} - \bar{\mathbf{X}}_{wd})\boldsymbol{\beta}_{md} + (\bar{\mathbf{Z}}_{md} - \bar{\mathbf{Z}}_{wd})\mathbf{u}_{md}, \\
U_d &= \bar{\mathbf{X}}_{wd}(\boldsymbol{\beta}_{md} - \boldsymbol{\beta}_{wd}) + \bar{\mathbf{Z}}_{wd}(\mathbf{u}_{md} - \mathbf{u}_{wd}).
\end{aligned}$$

and,

$$GPG_{Qd} = GPG_d\frac{Q_d}{\Delta_d} \qquad \text{and} \qquad GPG_{Ud} = GPG_d\frac{U_d}{\Delta_d}.$$

where

$$GPG_d = \frac{E(W_{md}) - E(W_{wd})}{E(W_{md})},$$

Small area estimators for Y_{gd}, Δ_d, U_d and Q_d (as in Rao and Molina, 2015) and estimators for $E(W_{gd}), g \in (m,w)$ $d = 1,...,D$, are defined as follows:

$$
\begin{array}{rcl}
\hat{\Delta}_d & = & \bar{\mathbf{X}}_{md}\hat{\boldsymbol{\beta}}_{md} + \bar{\mathbf{Z}}_{md}\hat{\mathbf{u}}_{md} - \bar{\mathbf{X}}_{wd}\hat{\boldsymbol{\beta}}_{wd} - \bar{\mathbf{Z}}_{wd}\hat{\mathbf{u}}_{wd}, \\
\hat{Q}_d & = & (\bar{\mathbf{X}}_{md} - \bar{\mathbf{X}}_{wd})\hat{\boldsymbol{\beta}}_{md} + (\bar{\mathbf{Z}}_{md} - \bar{\mathbf{Z}}_{wd})\hat{\mathbf{u}}_{md}, \\
\hat{U}_d & = & \hat{\Delta}_d - \hat{Q}_d.
\end{array} \tag{3}
$$

$$
\widehat{E(W_{gd})} = \overline{exp(\hat{Y}_{gd})}, g \in (m,w), d = 1,...,D; \tag{4}
$$

From (4), we have that:

$$
\widehat{GPG}_d = \frac{\widehat{E(W_{md})} - \widehat{E(W_{wd})}}{\widehat{E(W_{md})}}, \tag{5}
$$

Third Step: bias corrected estimators and confidence intervals.

Omitted variable bias is taken into account as in many applications, the implicit assumption that all confounding variables are included is not easy to attain. The bias in Q_d is derived by splitting the men sample into two parts I times. The value of B_d^i, for iteration $i = 1,...,I$, is expected to be zero when comparing the populations representing these two parts. A non-zero estimated value is assumed to be due to the omitted variables and is denoted by $\hat{B}_d^i$, for iteration i. From these values, the bias in Q_d is estimated, and bias-corrected estimators for GPG_{Qd} and GPG_{Ud} are derived, as explained below. Iterations with no men or women in any of the activities are discarded.

For each iteration $i = 1,\ldots,I$ repeat:

1. Randomly split the data S in two parts of size $n/2$: S_1^i and S_2^i. S_1^i of size $n/2$ is selected randomly in S so that $S_2^i = S - S_1^i$. In the following, subindex 1 or 2 indicates the data coming from S_1^i or S_2^i, respectively.

2. The model selected is fitted to men and women's data in S_1^i, and $\hat{Q}_{1d}^i$, $\hat{\Delta}_{1d}^i$, and $\widehat{GPG}_{1d}^i$ are calculated according to (3) and (5).

3. For each area d and iteration i, the bias is calculated considering men's data from S_1^i and S_2^i, as follows:
$$\hat{B}_d^i = (\bar{\mathbf{X}}_{1md} - \bar{\mathbf{X}}_{2md})\hat{\boldsymbol{\beta}}_{1md} + (\bar{\mathbf{Z}}_{1md} - \bar{\mathbf{Z}}_{2md})\hat{u}_{1md}.$$

Then, the final bias term is estimated as:

$$
\hat{B}_d = \frac{1}{I}\sum_{i=1}^{I}\hat{B}_d^i,
$$

and the bias corrected estimates of Q_d and U_d are:

$$
\hat{Q}_d^B = \hat{Q}_d + \hat{B}_d,
$$

$$
\hat{U}_d^B = \hat{\Delta}_d - \hat{Q}_d^B,
$$

where $\hat{Q}_d$ and $\hat{\Delta}_d$ are from the complete sample, as in (3).

Now, the bias corrected estimates of the explained and unexplained parts of GPG are derived from the latter and from (5) as:

$$\widehat{GPG}_{Qd} = \widehat{GPG}_d \frac{\hat{Q}_d^B}{\hat{\Delta}_d} \quad \text{and} \quad \widehat{GPG}_{Ud} = \widehat{GPG}_d \frac{\hat{U}_d^B}{\hat{\Delta}_d} \tag{6}$$

Moreover, confidence intervals are also derived as follows:
From S_1^i, $i = 1, \ldots, I$, and (4), we have that:

$$\hat{Q}_d^{Bi} = \hat{Q}_{1d}^i + \hat{B}_d, \quad \text{and} \quad \hat{U}_d^{Bi} = \hat{\Delta}_{1d}^i - \hat{Q}_d^{Bi}.$$

$$\widehat{GPG}_{1d}^i = \frac{E(\widehat{W_{1md}}) - E(\widehat{W_{1wd}})}{E(\widehat{W_{1md}})},$$

$$\widehat{GPG}_{Qd}^i = \widehat{GPG}_{1d}^i \frac{\hat{Q}_d^{Bi}}{\hat{\Delta}_{1d}^i},$$

$$\widehat{GPG}_{Ud}^i = \widehat{GPG}_{1d}^i \frac{\hat{U}_d^{Bi}}{\hat{\Delta}_{1d}^i},$$

Now, let us define:

$$\begin{aligned} V(\widehat{GPG}_{Qd}) &= \tfrac{1}{I} \textstyle\sum_{i=1}^I \left(\widehat{GPG}_{Qd}^i - \overline{\widehat{GPG}}_{Qd}^B \right)^2, \\ V(\widehat{GPG}_{Ud}) &= \tfrac{1}{I} \textstyle\sum_{i=1}^I \left(\widehat{GPG}_{Ud}^i - \overline{\widehat{GPG}}_{Ud}^B \right)^2, \end{aligned}$$

and,

$$\overline{\widehat{GPG}}_{Qd}^B = \frac{1}{I} \sum_{i=1}^I \widehat{GPG}_{Qd}^i.$$

$$\overline{\widehat{GPG}}_{Ud}^B = \frac{1}{I} \sum_{i=1}^I \widehat{GPG}_{Ud}^i.$$

while the confidence intervals are derived as follows:

$$\begin{aligned} &\widehat{GPG}_{Qd} \pm Z_{\alpha/2} \sqrt{V(\widehat{GPG}_{Qd})} \\ &\widehat{GPG}_{Ud} \pm Z_{\alpha/2} \sqrt{V(\widehat{GPG}_{Ud})} \end{aligned} \tag{7}$$

The aim of this document is to provide an easy-to-apply methodology for constructing confidence intervals to allow for inferences to be made. The confidence intervals provided are based on a Monte-Carlo variance estimation and on the assumption of normality (at least asymptotically) of the estimated decomposition of GDP proposed. The post-selection inference is an interesting issue that could be considered in the proposed methodology following the ideas of Berk et al. (2013), Lee et al. (2016) and Tibshirani et al. (2016), among others; but it is outside the scope of this document.

4 Estimation of the GPG by ecomomic activities in Galicia.

The data come from the Structure of Earnings Survey (SES) conducted by the Spanish National Institute of Statistics (INE) in 2014. The subset considering in this paper, corresponds to the Galicia region at the Northwest of Spain. The survey is conducted in a similar way to wage structure surveys in other European countries; it uses a two-level sampling of national companies, a stratified sampling in the first stage for local units and systematic sampling to select workers at those units. The economic activities are defined by the Statistical Classification of Economic Activities in the European Community (NACE09 Rev.2) at the division level.

Galicia has an economy strongly linked to natural resources, and its labor market has experienced growth in recent years. The GPG has been reduced from 15.5% in 2010 to 13.5% in 2016. This latter value is below that of the EU (16.3%), and below that of Spain (15.1%) in that year.

The survey collects information on non-self-employed workers who work in establishments with at least 10 employees and covers a wide range of private sectors (industry, construction, commerce, catering business, transport, financial intermediation, ...) excluding the primary sector. The total sample size in Galicia being 10276. The list of explanatory variables considered in this study are included in Table 1, the selection being similar to that in Aláez et al., (2000, 2001, 2003), De la Rica et al. (2008) or Moral-Arce et al. (2011).

The economic activities, which are defined by the Statistical Classification of Economic Activities in the European Community (NACE09 Rev.2) at the division level, are the areas of interest. Tables 10 and 11 in the Appendix include the list of codes and the description of the NACE09 at the division level, excluding the primary sector. In the present study, we have information for $D = 78$ of the 84 NACE09's. SAE methods are appealing because the sample sizes in some activities are very small. Specifically, the quartiles of the sample sizes distribution in the areas are $Q_1 = 26$, $Q_2 = 50$ and $Q_3 = 104$, for men, and $Q_1 = 15$, $Q_2 = 26$ and $Q_3 = 50$, for women.

Table 2 gives the percentage distribution for men and women regarding labor characteristics. The numbers in the table show a greater percentage of women in higher education levels, 41% against 33% of men, and that almost half the men (48%) are employed in the manufacturing industry or in public administration, while public administration (41%) and the retail trade (16%) are the activities with higher percentages for women. Regarding their type of work, women are more frequently on part-time work (27% vs 10% for men).

The distribution of the response variable $Y = log(W)$, where W is the hourly wages, is shown in Figure 1 for men and women; the shape of both curves being quite similar, though that of men is shifted to the right, which indicates an increase in salaries. Specifically, the quartiles are Q1 = 7.6, Q2 = 9.6 and Q3 = 13.6 for men and Q1 = 6.4, Q2 = 8.1 and Q3 = 12.1 for women. Finally, Figure 2 shows important differences in mean wages per hour across activities and between sexes. These differences range from 9 euros in the median of the activity *retail trade* to the maximum of 18 euros for those employed in *finance* and *insurance* activities.

The candidate list of models have been built as follows; First, the explanatory variables more often used in similar studies have been selected and have been described in Table 1. Next, the candidate models are defined combining fixed effect for Experience, Age, Education, Occupation, Contract, Type of work, Size of the enterprise, and Market with the effects by the Economic Activities. The latter can be incorporated in the model using fixed effects for aggregated levels ($X19$ to $X27$), or fixed effects for each activity ($X28$ to $X104$), or a random

Name	Variable	Description
Wage per hour	Y	Gross hourly earnings from employment
Gender		Men, Women
Experience	X_1	Number of years in the actual enterprise
Age	X_2	Years of the employee
Education		(Primary)
	X_3	Secondary
	X_4	Higher
Occupation		(Professionals and managers)
	X_5	Tecnic
	X_6	Operators
	X_7	Services and sales workers
	X_8	Non skilled workers
Contract	X_9	(Long term), Short term
Type of work	X_{10}	(Full time), Part time
Size of the enterprise		(Between 10 to 19 employees)
	X_{11}	Between 20 to 49 employees
	X_{12}	Between 50 to 99 employees
	X_{13}	Between 100 to 199 employees
	X_{14}	Between 200 to 499 employees
	X_{15}	More than 500 workers
Market		(Local or regional)
	X_{16}	National
	X_{17}	UE
	X_{18}	International
Aggregated economic activity		(Energy)
	X_{19}	Manufacturing industry
	X_{20}	Construction
	X_{21}	Retail trade
	X_{22}	Transportation, storage and Accommodation
	X_{23}	Information and Communication
	X_{24}	Finance and insurance
	X_{25}	Professionals
	X_{26}	Public Administration
	X_{27}	Other service
Economic activity for two digits	$X_{28} - X_{104}$	(7) NACE09 Rev. 2, division level, two digits

Table 1: Variables in the survey, in brackets the category used as a reference. For the variable *Economic activity for two digits*, the reference category is the activity 7, because in Galicia there are not people working in the activity 6 (see Appendix).

effect. Moreover, when a random effect is used to model the activities, the interactions with Experience, Education, and Occupation may also be considered.

Regarding the notation, u_d is the random effect for the Economic Activities, $u_{dj}, j = 1, 2$ for the interaction with Education, u_{dk} $k = 1, 2, 3, 4$ for the interaction with Occupation; and v_d for the interaction with Experience. Only a selection of the most interesting models is presented in Table 3 to simplify the exposition. Specifically, eight models labeled as $M1$ to $M8$ are presented,

	Men	Women
Experience (mean years)		
	11.16	10.43
Education (%)		
Primary	0.19	0.13
Secondary	0.48	0.46
Higher	0.33	0.41
Contract (%)		
Long-term	0.77	0.77
Short-term	0.23	0.23
Type of work (%)		
Full time	0.90	0.73
Part time	0.10	0.27
Economic activity (%)		
Energy	0.03	0.01
Manufacturing industry	0.27	0.13
Construction	0.09	0.01
Retail trade	0.13	0.16
Transportation and storage and Accommodation	0.08	0.07
Information and Communication	0.03	0.02
Finance and insurance	0.03	0.03
Professionals	0.10	0.14
Public Administration	0.21	0.41
Other services	0.02	0.02

Table 2: Summary of labor force characteristics of the people in the survey by gender.

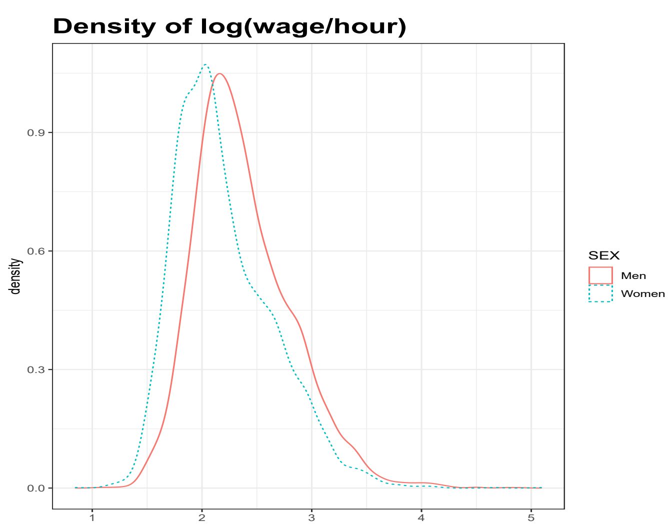

Figure 1: Density of log(wage/hour) by sex in Galicia.

models $M6$ and $M7$ are defined with fixed effects, and the rest with a combination of fixed and random effects. Models defined by other combinations of effects have been considered but have

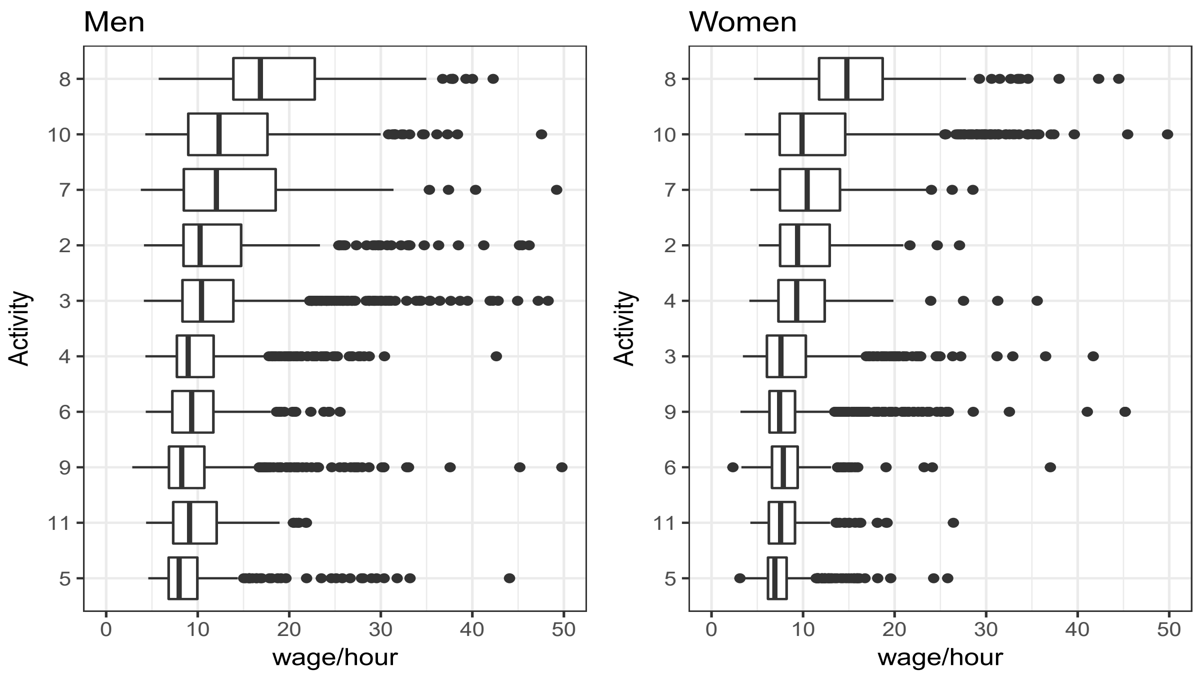

Figure 2: Wage per hour by sex and economic activity in Galicia. The codes for the activities are: 2 Energy, 3 Manufacturing industry, 4 Construction, 5 Retail trade, 6 Transportation and storage and Accommodation, 7 Information and Communication, 8 Finance and insurance, 9 Professionals, 10 Public Administration and 11 Other services.

been discarded because the results are far from being better than those obtained with other models.

Label	Fixed effects	Random Effects	Fixed parameters (nº)	Random parameters (nº)
$M1$	X_1-X_{18}	u_d	18	1
$M2$	X_1-X_{18}	v_d	18	1
$M3$	X_1-X_{18}	u_d, v_d	18	2
$M4$	X_1-X_{18}	u_{dj}	18	2
$M5$	X_1-X_{18}	u_{dk}	18	4
$M6$	X_1-X_{27}	-	27	0
$M7$	X_1-X_{18}, X_{28}-X_{104}	-	95	0
$M8$	X_1-X_{18}, X_{28}-X_{104}	v_d	95	1

Table 3: Models used to fit the data.

In order to explore differences between models, the estimated bias for each activity has been calculated for each candidate model and is shown in Figure 3. The activities are sorted in increasing order of bias according to the $M1$ model. Figure 3 shows important bias in some activities: 88 = S*ocial work activities without accommodation*, 90 = *Creative, arts and entertainment activities* or 39 = *Remediation activities and other waste management services*, among others. Moreover, the bias corrected estimator of the explained and unexplained parts of the GPG for each economic activity and model, as shown in the expression (6), are shown in Figures 4 and 5, respectively, sorting the activities by the values of $\widehat{GPG}_{Qd}$ and $\widehat{GPG}_{Ud}$ for

$M1$. These Figures show that while estimates for $\widehat{GPG}_{Qd}$ are similar between models, there are important differences in $\widehat{GPG}_{Ud}$ across models.

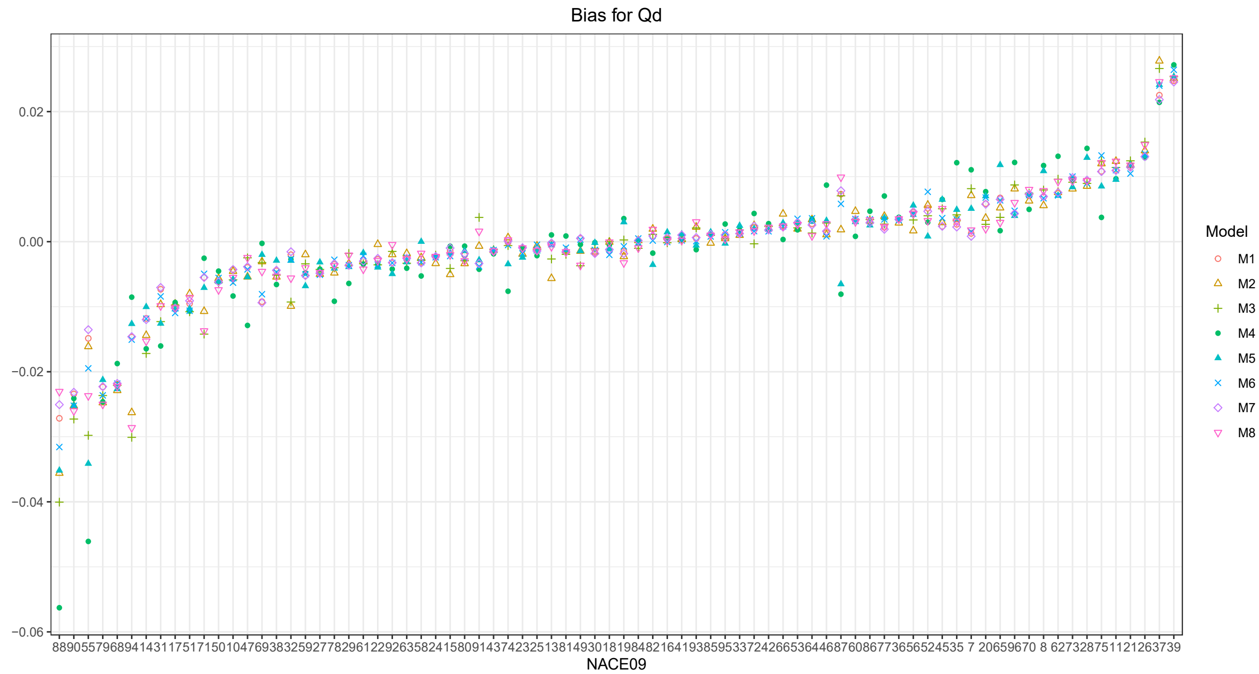

Figure 3: Bias for Q_d ($\hat{B}_d$), for each activity d defined by the NACE09.

The model selected, using the proposal described in Section 3, is $M1$. The parameter estimates for $M1$ are shown in Table 4, which explains that the wage of a worker increase with higher education, working in an enterprise with a national, UE or international market, and belonging to a large enterprise. The coefficients related to occupations show negative signs as *Professionals and managers* is the reference category. The part-time work variable has been included in the model despite not being significant due to the importance given to it by the experts. The results are consistent with those from other studies (Aláez et al. (2003), De la Rica (2008) and Moral-Arce et al. (2011)).

Pay gap estimators have been obtained using the model $M1$ and the proposal in Section 3. In Galicia, the pay gap between men and women ($\widehat{GPG}$) is estimated to be of 11.3%, where 2.8% is due to the observed quantify effect ($\widehat{GPG}_Q$) and 8.5% to discrimination ($\widehat{GPG}_U$), as Table 5 shows.

Moreover, Figures 6 and 7 show $\widehat{GPG}_{Qd}$ and $\widehat{GPG}_{Ud}$, respectively, for each activity, d, with their confidence intervals. Activities are ordered by the increasing value of the estimator.

It is worth commenting on the results from activities 84 and 85, which correspond to the Public Administration and Education. They are useful for validating the procedure because no discrimination is expected in these cases as the Administration control the wages (Aláez et al., 2000, Moral-Arce et al., 2011). Figure 6 shows no significant differences between the characteristics of men and women, for activities 84 and 85; while Figure 7 shows, for 84 and 85, no significative differences from those of Galicia. The results are consistent with what would be

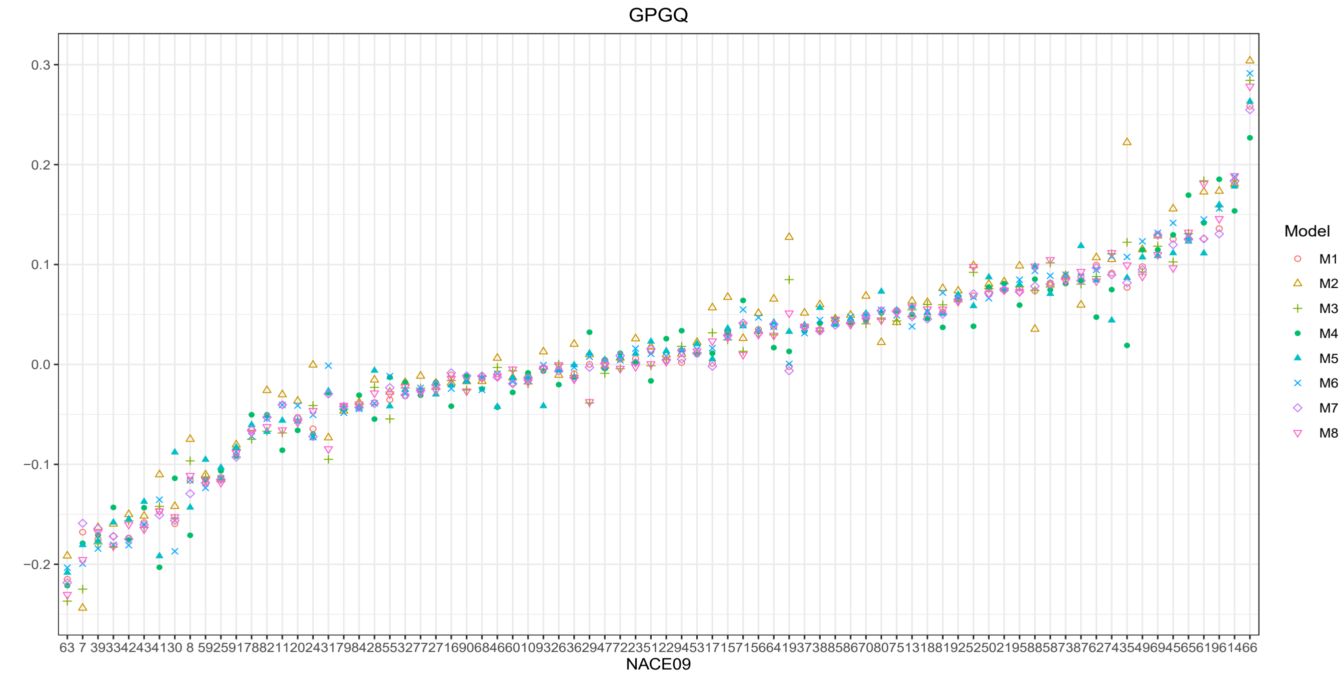

Figure 4: $\widehat{GPG}_{Qd}$ for all the models considered and for each activity d defined by the NACE09.

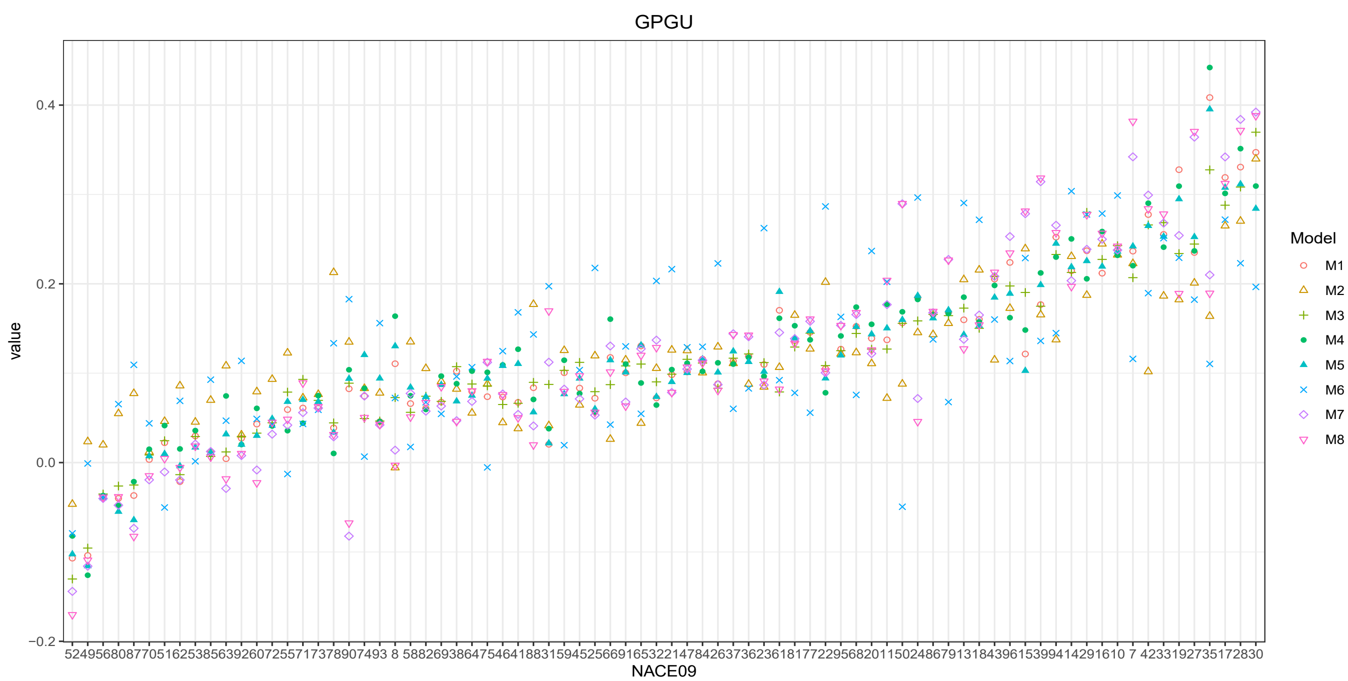

Figure 5: $\widehat{GPG}_{Ud}$ for all the models considered and for each activity d defined by the NACE09.

expected.

On the other hand, the results show very high discrimination in eight activities: 30, 28,

	Estimate	Std. Error	p-value
Fixed effects			
(Intercept)	2.299	0.033	0.000
Experience (X_1)	0.007	0.001	0.000
Age (X_2)	0.005	0.000	0.000
Secondary (X_3)	0.015	0.011	0.188
Higher (X_4)	0.122	0.014	0.000
Technic (X_5)	-0.248	0.015	0.000
Operators (X_6)	-0.439	0.016	0.000
Services and sales workers (X_7)	-0.504	0.019	0.000
Non-skilled workers (X_8)	-0.527	0.016	0.000
Short term contract (X_9)	-0.098	0.010	0.000
Part time work (X_{10})	-0.017	0.013	0.210
Between 20 to 49 workers (X_{11})	-0.029	0.012	0.028
Between 50 to 99 workers (X_{12})	0.047	0.014	0.003
Between 100 to 199 workers (X_{13})	0.042	0.014	0.011
Between 200 to 499 workers (X_{14})	0.081	0.014	0.000
More than 500 workers (X_{15})	0.113	0.016	0.000
National Market (X_{17})	0.052	0.010	0.000
UE Market (X_{18})	0.076	0.018	0.001
International Market (X_{19})	0.072	0.016	0.000
Random effects			
σ_u^2	0.021		

Table 4: Coefficient estimates for men and for $M1$.

	Estimation	CI 95%
$\widehat{GPG_Q}$	0.028	(-0.018,0.07)
$\widehat{GPG_U}$	0.085	(0.057,0.14)
$\widehat{GPG}$	0.113	-

Table 5: Decomposition of gender pay gap in Galicia.

19, 17, 42 , 33, 29 and 10, almost all of them are from the industrial sector (see Figure 7); while differences between the characteristics due to gender are not significantly higher for these activities than for Galicia (Figure 6).

Taking the GPG obtained for the region of Galicia as reference, $\widehat{GPG} = 11.3\%$, Table 9 shows the activities with $\widehat{GPG_d}$ greater than 11.3%, sorted from highest to lowest. The activity at the top is *Electricity, gas, steam and air conditioning supply* and it is also one of the activities where there are more discrimination. The second activity at the top is *Manufacture of wearing apparel*, and this is an example that high wage differences are not necessarily accompanied by discrimination. This activity is very important in the Galician business sector and employs approximately 8,600 workers, 1% of the total number of employed people in Galicia. On the other hand, note that only ten activities out of 74 have positive discrimination significantly higher than the region's mean. These activities employ 8.2% of the workers in the region. It is also interesting to note that there are two activities in which discrimination is significantly negative (in favor of women). These are *Warehousing and support activities for transportation*

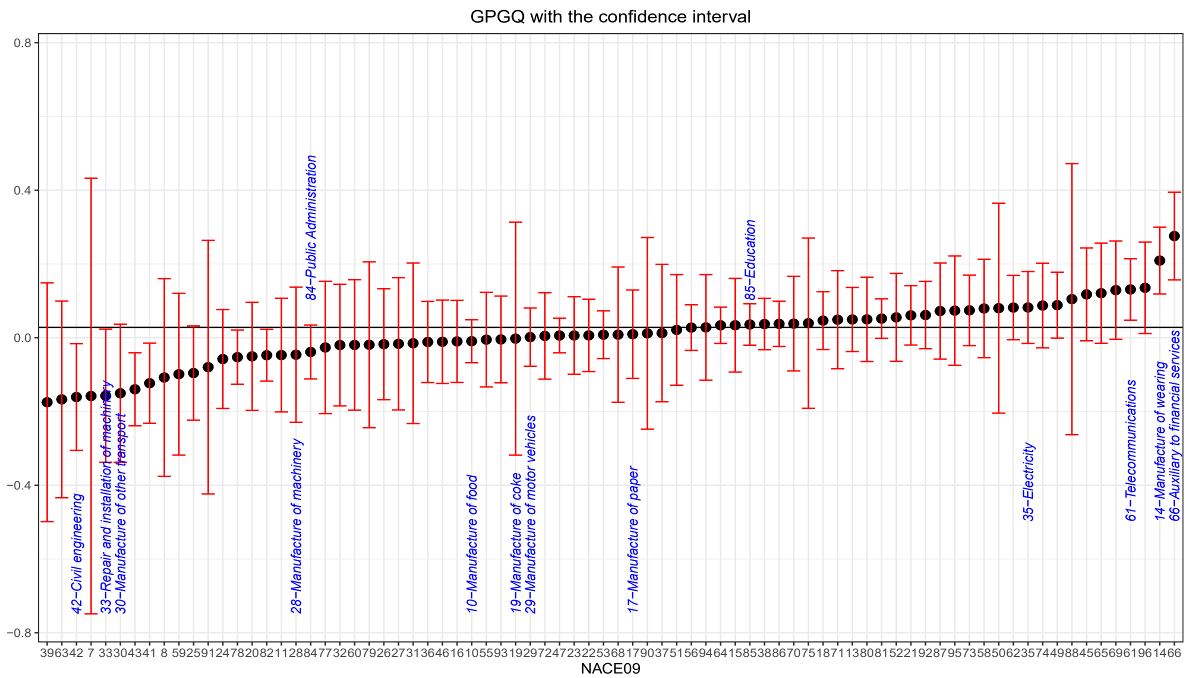

Figure 6: $\widehat{GPG}_{Qd}$ with their confidence interval (95%) and for each activity d defined by the NACE09. The horizontal line is the $\widehat{GPG}_Q$ for Galicia.

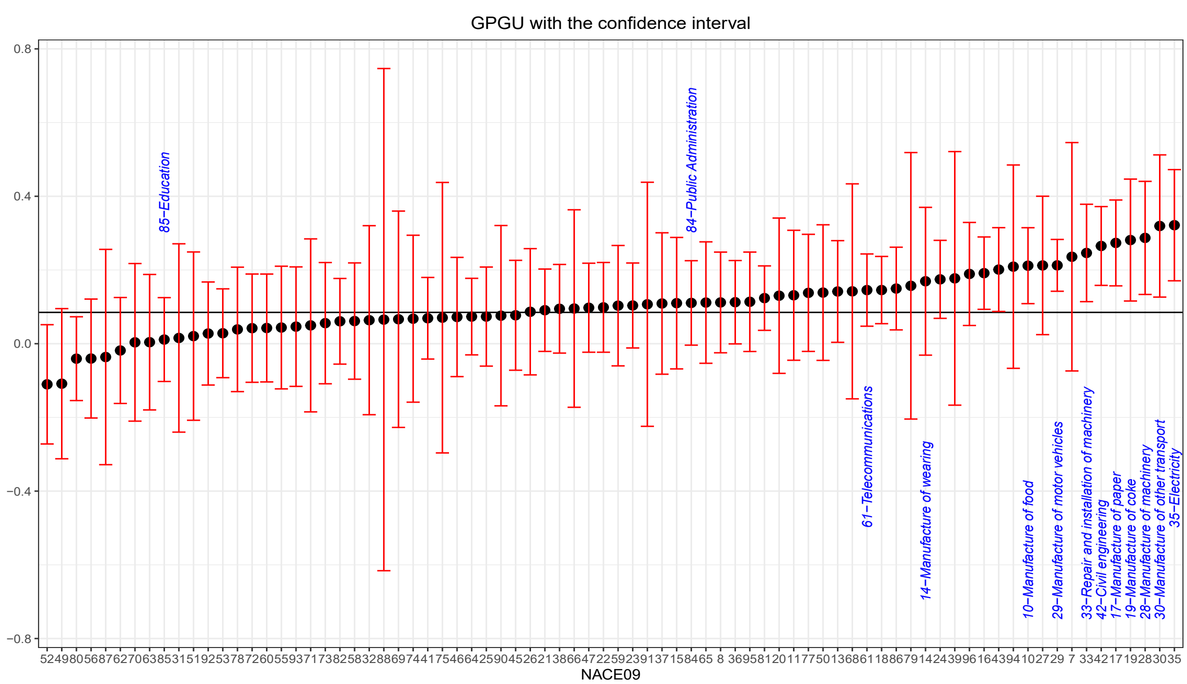

Figure 7: Discrimination ($\widehat{GPG}_{Ud}$) with their confidence interval (95%) and for each activity d defined by the NACE09. The horizontal line is the $\widehat{GPG}_U$ for Galicia.

and *Security and investigation activities* that belong to the service sector where women have a more significant presence than men.

5 Simulations

The aim of this section is to validate the novel proposal and compare it with the classical approach of OB. Moreover, the estimation error that results when important explanatory variables are omitted, is also illustrated.

Two simulation experiments are constructed. The first simulation experiment imitates SES real data and it is based in a model more complex than any of the models in the candidate list. The generating model includes a fixed parameter for each activity and explicative variables; being the candidate models much more simple. In the second simulation experiment the data is generated with M1, the model selected in the real case.

In both simulation experiments the 30 activities which have the largest sample sizes and the explanatory variables in Table 1 are considered. The sample sizes take values from 21 to 111 for men and 7 to 82 for women.

A detailed description of the simualtion experiment designs is given below:

- Simulation 1

  $$Y_{gdi} = X_{1gdi}\beta_{1gd} + X_{3gdi}\beta_{2gd} + X_{4dg}\beta_{3gd} + X_{5gdi}\beta_{4gd} + X_{6gdi}\beta_{5gd} + X_{7gdi}\beta_{6gd} + X_{8gdi}\beta_{7gd} + \epsilon_{gdi}$$

  where $g \in \{m, w\}$, $d = 1, ..., 30$, $i = 1, \ldots, n_{gd}$, $\epsilon_{gdi} \sim N(0, \sigma_{gd}^2)$, with $\sigma_{gd} = 0.1$ and β_{jgd}, $j = 1, \ldots, 7$ are the estimated values using model $M6$.

  In this simulation we considerate two scenarios, when the model is estimated with all the variables (scenario 1.1) and when variables X_3 and X_4 are omitted (scenario 1.2).

- Simulation 2

  $$Y_{gdi} = X_{1gdi}\beta_{1g} + X_{3gd}\beta_{2g} + X_{4gdi}\beta_{3g} + X_{5gdi}\beta_{4g} + X_{6gdi}\beta_{5g} + X_{7gdi}\beta_{6g} + X_{8gdi}\beta_{7g} + u_{gd} + \epsilon_{gdi}$$

  where $g \in \{m, w\}$, $d = 1, ..., 30$, $i = 1, \ldots, n_{gd}$, $\epsilon_{gdi} \sim N(0, \sigma_{gd}^2)$, with $\sigma_{gd} = 0.3$ and $u_{gd} \sim N(0, \sigma_{ug}^2)$, and β_{jg}, $j = 1, \ldots, 7$.

  In this simulation we considerate three scenarios:

  - Scenario 2.1: $\sigma_{ug} = 0.15$ (real case).
  - Scenario 2.2: $\sigma_{ug} = 0.3$.
  - Scenario 2.3: $\sigma_{ug} = 0.45$.

For each d, GPG_d, GPG_{Qd} and GPG_{Ud} are obtained using (1), (2) and (5) and are assumed to be true values.

Label	Fixed effects	Random Effects	Fixed parameters (nº)	Random parameters (nº)
$MS1$	X_1,X_3-X_8	u_d	7	1
$MS2$	X_1,X_3-X_8	v_d	7	1
$MS3$	X_1,X_3-X_8	u_d, v_d	7	2
$MS4$	X_1,X_3-X_8	u_{dj}	7	2
$MS5$	X_1,X_3-X_8	u_{dk}	7	4
$MS6$	X_1,X_3-X_8, X_{19}-X_{27}	-	16	0
$MS7$	X_1,X_3-X_8, X_{28}-X_{56}	-	36	0
$MS8$	X_1,X_3-X_8, X_{28}-X_{56}	v_d	36	1

Table 6: Models used to fit the data.

The models in Table 6 combine fixed and random effects, as in the real case. Estimators for GPG components and each model are obtained using proposal in Section 3.

Now, the empirical mean square error (EMSE) of $\widehat{GPG}_{Qd}$ and $\widehat{GPG}_{Ud}$ for all the models and for each activity d are as follows:

$$EMSE(\widehat{GPG}_{Qd}) = \frac{1}{500}\sum_{i=1}^{500}\left(\widehat{GPG}_{Qd}^{(i)} - GPG_{Qd}\right)^2$$

$$EMSE(\widehat{GPG}_{Ud}) = \frac{1}{500}\sum_{i=1}^{500}\left(\widehat{GPG}_{Ud}^{(i)} - GPG_{Ud}\right)^2$$

where $\widehat{GPG}_{Qd}^{(i)}$ and $\widehat{GPG}_{Ud}^{(i)}$ are the estimations for each iteration $i = 1, \dots, 500$.

Finally, we calculate the confidence intervals for GPG_{Qd} and GPG_{Ud} from (7) and give the percentage of times that the true values GPG_{Qd} and GPG_{Ud} are in the corresponding confidence interval.

Tables 7 and 8 present the results of the simulations. Table 7 shows the mean of the EMSE of $\widehat{GPG}_{Qd}$ and $\widehat{GPG}_{Ud}$ obtained from the models $MS1$ to $MS8$, the OB model and the model selected by xGAIC (XG) (in rows) for the different scenarios (in columns). Table 8 presents the same information but for the coverage rates of confidence intervals.

In Table 7 we can see that the best models, in terms of EMSE, are $MS5$ and XG for Simulation 1 and $MS1$ and XG for Simulation 2, as expected. The error increases in scenario 1.2 relative to scenario 1.1, more striking when OB is used and for the model MS6. The conclusions are similar regarding the coverage rates of 95% confidence intervals in Table 8. It is interesting to see how the coverage of the confidence interval from scenario 2.1 to scenario 2.3 is close to 95% in all the cases, except for the OB, MS2 and MS6 models, which none of them consider the random effect of the activity. In summary, in Tables 7 and 8, you can see that the models selected by the $xGAIC$, XG models, have a good behavior, because they are among those that have a lower EMSE and a greater coverage. For example, for Simulation 1, when the model is fitted with all the variables, the $xGAIC$ selects the MS5 model 82% of times, 15% of times $MS4$ and 3% $MS8$. When X_3 and X_4 are simultaneously removed from the working model (scenario 1.2), the $xGAIC$ selects the model $MS5$ 98% of times and 2% the model $MS8$.

Finally, this study highlights the gain of our methodology with respect to the OB methodology, especially for estimating GPG_{Ud}.

	Scenario									
	1.1		1.2		2.1		2.2		2.3	
Model	GPG_Q	GPG_U	GPG_Q	GPG_U	GPG_Q	GPG_U	GPG_Q	GPG_U	GPG_Q	GPG_U
OB	0.0050	0.0067	0.0059	0.0089	0.0159	0.0389	0.0179	0.1440	0.0073	0.4079
MS1	0.0046	0.0044	0.0051	0.0056	0.0159	0.0178	0.0175	0.0217	0.0065	0.0113
MS2	0.0057	0.0065	0.0064	0.0074	0.0163	0.0281	0.0195	0.0758	0.0116	0.2063
MS3	0.0060	0.0047	0.0063	0.0056	0.0159	0.0182	0.0175	0.0217	0.0065	0.0113
MS4	0.0047	0.0038	0.0051	0.0056	0.0159	0.0179	0.0176	0.0217	0.0066	0.0114
MS5	0.0045	0.0042	0.0051	0.0053	0.0161	0.0183	0.0178	0.0219	0.0067	0.0116
MS6	0.0050	0.0067	0.0059	0.0079	0.0159	0.0386	0.0179	0.1428	0.0073	0.4077
MS7	0.0046	0.0046	0.0051	0.0059	0.0159	0.0189	0.0175	0.0222	0.0066	0.0116
MS8	0.0060	0.0048	0.0063	0.0058	0.0159	0.0190	0.0175	0.0222	0.0066	0.0116
XG	0.0046	0.0043	0.0051	0.0056	0.0159	0.0178	0.0175	0.0217	0.0066	0.0113

Table 7: EMSE of GPG_Q and GPG_U for the model in Simulation 1 and 2. OB is the Oaxaca-Blinder model and XG is the model selected by the $xGAIC$

	Scenario									
	1.1		1.2		2.1		2.2		2.3	
Model	GPG_Q	GPG_U	GPG_Q	GPG_U	GPG_Q	GPG_U	GPG_Q	GPG_U	GPG_Q	GPG_U
OB	86.0	54.6	80.6	43.8	94.4	45.3	96.5	31.3	95.5	26.1
MS1	86.2	78.3	81.1	73.1	94.6	92.8	97.1	96.4	96.8	96.8
MS2	83.1	57.7	79.5	54.1	93.8	55.2	94.5	49.6	92.5	45.1
MS3	89.0	81.1	85.2	77.1	94.6	92.3	97.1	96.4	96.8	96.9
MS4	89.7	82.4	81.1	73.1	94.6	92.9	97.1	96.5	96.9	96.8
MS5	90.0	84.0	86.5	80.8	94.8	93.0	97.3	96.6	97.0	97.0
MS6	86.0	54.6	80.6	45.8	94.4	45.6	96.5	31.3	95.5	26.3
MS7	86.1	81.1	80.9	75.7	94.6	95.1	97.1	97.1	96.8	97.0
MS8	88.9	82.9	85.0	78.7	94.7	95.1	97.1	97.1	96.8	97.1
XG	89.9	83.7	86.5	80.8	94.6	93.0	97.1	96.4	97.0	96.8

Table 8: Coverage of the GPG_Q and GPG_U confidence intervals for the model in Simulation 1 and 2. OB is the Oaxaca-Blinder model and XG is the model selected by the $xGAIC$

6 Conclusions

The first contribution of this paper, and the most important from the theoretical point of view, is the derivation of a novel approach to estimate the components of the gender pay gap in small areas. This methodology is consistent with that used by Eurostat (Leythienne and Ronkowski, 2018), it also generates confidence intervals for the explained and unexplained part of the GPG, includes a bias correction and considers linear mixed models and a selection of models option. Moreover, this novel approach is robust against potential differences in unobserved variables.

The second, and most important contribution from the practical point of view, is that the application of the methodology to estimate wage discrimination in economic activities in Galicia reveals important differences in the region and among activities.

In Galicia, 25% of the pay gap is explained by the differences between characteristics and 75% is due to discrimination. In other words, if there were no discrimination and the characteristics of men and women had been paid at the same prices, the wage differential would be reduced to

2.8%.

Regarding the activities, a high pay gap due to differences in the characteristics is shown in such activities as 66 and 14, while a high pay gap due to discrimination is shown in other activities, such as 28, 19, 17, among others.

The existence of salary differences by sex continues to be a real problem, according to the results obtained. The *GPG* indicator, together with the explained and unexplained gap by economic activities, allow for a better identification and interpretation of the causes of the gender pay gap. As a consequence, policy actions can be better targeted.

Finally, to determine whether two workers have the same qualification and value, useful information, such as the total work experience, knowledge of languages or the awards or publications obtained, among many others, is still missing. The inclusion of this information as an explanatory variable would likely derived in more accurate pay gap estimators. Indeed, the selection of an appropriate set of confounders is essential for reliable causal inference. GAIC is a method with interesting advantages that is widely used in the statistical literature on variable selection. However, a more rigorous study on the selection of confounders and their impact on causal inference is needed. We leave this topic open for future studies.

7 Appendix

Tables 10 and 11 show the codes and the description of the Statistical Classification of Economic Activities in the European Community (NACE09 Rev.2) at the division level, excluding the primary sector.

8 Bibliography

Akaike, H. (1973). Information theory and the maximum likelihood principle. In *International Symposium on Information Theory*, pages 267-281, Budapest. Akademiai Kiado.

Aláez, R., Ullibarri, M. (2000). Discriminación salarial por sexo: un análisis del sector privado y sus diferencia regionales en España. ICE, *Revista de Economía*, 789, 117-138.

Aláez, R., Longás, J.C., Ullibarri, M. (2001). Visualising gender wage differences in the European Union. *Gender, Work and Organization.*

Aláez, R., Longás, J.C., Ullibarri, M. (2003). Diferencias salariales en España: un análisis sectorial/regional. *Investigaciones regionales*, 3, 5-24.

Anastasiade, M.C. and Tillé, Y. (2017a). Decomposition of gender wage inequalities through calibration: Application to the Swiss structure of earnings survey. *Survey Metodology*, 43:211-234.

Anastasiade, M.C. and Tillé, Y. (2017b). Gender wage inequalities in Switzerland: The public versus the private sector. *Statistical Methods and Applications*, 26 (2) 293-316.

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401),28-36.

Berk, R., Brown, L. Buja, a. Zhang, K. and Zaho L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2), 802-837.

Blinder, A.S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4), 436-455.

De la Rica, S., Dolado, J.J., Llorens, V. (2008). Ceilings or floors: Gender wage gaps by education in Spain. *J. Popul Econ*, 21, 751-778

Dempster,A. P., Rubin D. B. and Tsutakawa R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341-353.

Fernández, M., Montuenga, V. M., Romeu, A E. (2000). Wage differencials and non competitive behaviour in the spanish industry labor market. *Documentos de traballo análise económica*, n 9, IDEGA.

Fan, Y., and Li, R. (2012). Variable selection in linear mixed effects models. *Annals of statistics*, 40(4), 2043-2068.

Fortin, N., Lemieux, T., and Firpo, S. (2011). Decomposition methods in economics. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 4, 1-102. Elsevier.

Graf, M. , Marín, M. and Molina, I. (2019). A generalized mixed model for skewed distributions applied to small area estimation. *TEST*, 28, 565-597.

Hlavac, Marek (2018). oaxaca: Blinder-Oaxaca decomposition in R. R package version 0.1.4. https://CRAN.R-project.org/package=oaxaca.

Jann B (2008). The Blinder-Oaxaca decomposition for linear regression models. *Stata Journal*, 8(4), 453-479.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Series in Statistics.

Juhn, C., Murphy, K. M., and Pierce, B. (1993). Wage inequality and the rise in returns to skill. *Journal of Political Economy*, 101(3):410-442.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor J. E. (2016). Exact post-selection inference with application to the lasso. *The Annals of Statistics*, 44(3), 907-927

Leythienne, D. abd Ronkowski P. (2018). A decomposition of the unadjusted gender pay gap using Structure of Earnings Survey data. *Population and social conditions. Statistical Working Papers*. Publications Office of the European Union.

Lombardía, M.J., López-Vizcaíno, E., and Rueda, C. (2017). Mixed generalized Akaike information (xGAIC) for small area models. *Journal of Royal Statistical Society Series A*, 180, 1229-1252.

Moral-Arce, I., Sperlich, S., Fernandez-Sáenz, A.I., Roca, M.J. (2011). Trends in the gender pay gap in Spain: A semiparametric analysis. *J. Labor Res*, 33, 173?195.

Oaxaca R.L. (1973). Male-female wage differentials in urban labor markets.*International Economic Review*, 14(3), 693-709.

Oaxaca R.L., Ransom M.R. (1994). On the discrimination and the decomposition of wage differentials. *Journal of Econometrics*, 61, 5-21.

Oaxaca R.L., Ransom M.R. (1998). Calculation of approximate variances for wage decomposition differentials. *Journal of Economic and Social Measurement*, 81(1), 154-157.

Opsomer, J., Claeskens, G., Ranalli, M., Kauermann, G., and Breidt, F. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of Royal Statistical Society Series B*, 70, 265-286.

Popli, G.K. (2013). Gender wage differentials in Mexico: a distributional approach. *Journal of the Royal Statistical Society Series A*, 176(2), 295-319.

Rao, J.N.K. and Molina, I. (2015). *Small area estimation*, Second Edition. J. Wiley, Hoboken (NJ).

Rueda, C. and Lombardía, M.J. (2012). Small area semiparametric additive monotone models. *Statistical Modelling*, 12, 527-549.

Tibshirani, R.J., Taylor J. Lockhart R. and Tibshirani R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514), 600-620.

NACE09	$\widehat{GPG}_d$	$\widehat{GPG}_{Qd}$	$\widehat{GPG}_{Qd}^-$	$\widehat{GPG}_{Qd}^+$	$\widehat{GPG}_{Ud}$	$\widehat{GPG}_{Ud}^-$	$\widehat{GPG}_{Ud}^+$
35	0.40	0.08	-0.02	0.18	0.32	0.17	0.47
14	0.35	0.21	0.12	0.30	0.17	-0.03	0.37
96	0.32	0.14	0.01	0.26	0.19	0.05	0.33
19	0.28	-0.00	-0.32	0.31	0.28	0.12	0.45
17	0.27	0.01	-0.11	0.13	0.27	0.16	0.39
61	0.27	0.13	0.05	0.21	0.15	0.05	0.24
28	0.25	-0.05	-0.23	0.14	0.29	0.13	0.44
65	0.24	0.12	-0.02	0.26	0.11	-0.05	0.28
94	0.22	0.03	-0.12	0.17	0.21	-0.07	0.48
50	0.21	0.08	-0.20	0.36	0.14	-0.05	0.32
29	0.21	0.00	-0.08	0.08	0.21	0.14	0.28
45	0.20	0.12	-0.01	0.24	0.08	-0.07	0.23
10	0.20	-0.01	-0.07	0.05	0.21	0.11	0.31
13	0.19	0.05	-0.04	0.14	0.14	0.00	0.28
18	0.19	0.05	-0.03	0.12	0.15	0.05	0.24
69	0.19	0.13	-0.00	0.26	0.07	-0.23	0.36
86	0.19	0.04	-0.02	0.10	0.15	0.04	0.26
27	0.19	-0.02	-0.20	0.16	0.21	0.02	0.40
95	0.19	0.07	-0.07	0.22	0.11	-0.02	0.25
16	0.18	-0.01	-0.12	0.10	0.19	0.09	0.29
81	0.18	0.05	-0.00	0.11	0.12	0.04	0.21
21	0.17	0.06	-0.02	0.14	0.09	-0.02	0.20
74	0.16	0.09	-0.03	0.20	0.07	-0.16	0.29
30	0.16	-0.15	-0.34	0.04	0.32	0.13	0.51
88	0.15	0.10	-0.26	0.47	0.07	-0.62	0.75
15	0.14	0.03	-0.09	0.16	0.11	-0.07	0.29
73	0.14	0.07	-0.02	0.17	0.06	-0.11	0.22
37	0.14	0.01	-0.17	0.20	0.11	-0.08	0.30
58	0.14	0.08	-0.05	0.21	0.06	-0.10	0.22
68	0.13	0.01	-0.18	0.19	0.14	-0.15	0.43
38	0.13	0.04	-0.03	0.11	0.09	-0.03	0.22
75	0.12	0.04	-0.19	0.27	0.07	-0.30	0.44
79	0.12	-0.02	-0.24	0.21	0.16	-0.20	0.52

Table 9: Observed gap, $\widehat{GPG}_{Qd}^B$ and $\widehat{GPG}_{Ud}^B$ with the confidence interval for the 33 activities with the highest observed gaps.

NACE09	Description
6	Extraction of crude petroleum and natural gas
7	Mining of metal ores
8	Other mining and quarrying
9	Mining support service activities
10	Manufacture of food products
11	Manufacture of beverages
12	Manufacture of tobacco products
13	Manufacture of textiles
14	Manufacture of wearing apparel
15	Manufacture of leather and related products
16	Manufacture of wood and of products of wood and cork, except furniture
17	Manufacture of paper and paper products
18	Printing and reproduction of recorded media
19	Manufacture of coke and refined petroleum products
20	Manufacture of chemicals and chemical products
21	Manufacture of basic pharmaceutical products and pharmaceutical preparations
22	Manufacture of rubber and plastic products
23	Manufacture of other non-metallic mineral products
24	Manufacture of basic metals
25	Manufacture of fabricated metal products, except machinery and equipment
26	Manufacture of computer, electronic and optical products
27	Manufacture of electrical equipment
28	Manufacture of machinery and equipment n.e.c.
29	Manufacture of motor vehicles, trailers and semi-trailers
30	Manufacture of other transport equipment
31	Manufacture of furniture
32	Other manufacturing
33	Repair and installation of machinery and equipment
35	Electricity, gas, steam and air conditioning supply
36	Water collection, treatment and supply
37	Sewerage
38	Waste collection, treatment and disposal activities; materials recovery
39	Remediation activities and other waste management services
41	Construction of buildings
42	Civil engineering
43	Specialised construction activities
45	Wholesale and retail trade and repair of motor vehicles and motorcycles
46	Wholesale trade, except of motor vehicles and motorcycles
47	Retail trade, except of motor vehicles and motorcycles
49	Land transport and transport via pipelines
50	Water transport
51	Air transport
52	Warehousing and support activities for transportation
53	Postal and courier activities
55	Accommodation
56	Food and beverage service activities

Table 10: List of codes for the NACE09 at division level, excluding the primary sector.

NACE09	Description
58	Publishing activities
59	Motion picture, video and television programme production, sound recording and music
60	Programming and broadcasting activities
61	Telecommunications
62	Computer programming, consultancy and related activities
63	Information service activities
64	Financial service activities, except insurance and pension funding
65	Insurance, reinsurance and pension funding, except compulsory social security
66	Activities auxiliary to financial services and insurance activities
68	Real estate activities
69	Legal and accounting activities
70	Activities of head offices; management consultancy activities
71	Architectural and engineering activities; technical testing and analysis
72	Scientific research and development
73	Advertising and market research
74	Other professional, scientific and technical activities
75	Veterinary activities
77	Rental and leasing activities
78	Employment activities
79	Travel agency, tour operator and other reservation service and related activities
80	Security and investigation activities
81	Services to buildings and landscape activities
82	Office administrative, office support and other business support activities
84	Public administration and defence; compulsory social security
85	Education
86	Human health activities
87	Residential care activities
88	Social work activities without accommodation
90	Creative, arts and entertainment activities
91	Libraries, archives, museums and other cultural activities
92	Gambling and betting activities
93	Sports activities and amusement and recreation activities
94	Activities of membership organisations
95	Repair of computers and personal and household goods
96	Other personal service activities
97	Activities of households as employers of domestic personnel
98	Undifferentiated goods- and services-producing activities of private households for own use
99	Activities of extraterritorial organisations and bodies

Table 11: List of codes for the NACE09 at division level, excluding the primary sector (continuation).