INSTITUTE OF ECONOMICS
Centre for Economic and Regional Studies

KRTK KTI

Közgazdaság- és Regionális Tudományi Kutatóközpont
KÖZGAZDASÁG-TUDOMÁNYI INTÉZET

CERS-IE WORKING PAPERS | KRTK-KTI MŰHELYTANULMÁNYOK

The gender-dependent structure of wages in Hungary: results using machine learning techniques

OLGA TAKÁCS – JÁNOS VINCZE

CERS-IE WP – 2020/44

November 2020

https://www.mtakti.hu/wp-content/uploads/2020/11/CERSIEWP202044.pdf

INSTITUTE OF ECONOMICS, CENTRE FOR ECONOMIC AND REGIONAL STUDIES,
BUDAPEST, 2020

ABSTRACT

This paper reports the results of a Blinder-Oaxaca style decomposition analysis on Hungarian matched employer-employee data to study the gender pay-gap. We carry out the decomposition by Random Forest regressions. The raw gap in our horizon (2008-2016) is increasing, but we find that the wage structure effects are rather stable, thus the rise in the gap is due to the disappearance of the formerly negative composition effects. Graphical analysis sheds light on interesting non-linear relationships; some of them can be readily interpreted by the previous literature. A Classification and Regression Tree analysis suggests that complicated interaction patterns exist in the data. We identify segments of the Hungarian labour market that are most and least exposed to gender-dependent wage determination. Our findings lend support to the idea that an important part of the gender wage gap is attributable to monopsonistic competition with gender-dependent supply elasticities.

JEL codes: J7 J3 C14
Keywords: Gender pay gap
Blinder-Oaxaca decomposition
Random Forest Regression
Hungary

Olga Takács
Corvinus University of Budapest, Hungary, H-1093 Budapest, Fővám Square 8
e-mail: olga.takacs@stud.uni-corvinus.hu

János Vincze
Corvinus University of Budapest, Hungary , H-1093 Budapest, Fővám Square 8
and
Centre for Economic and Regional Studies, Institute of Economics, (KRTK KTI), Hungary, H-1097, Budapest, 4 Tóth Kálmán street.
e-mail: Vincze.Janos@krtk.mta.hu

**

Nemtől függő bérstruktúra hatások Magyarországon: becslés gépi tanulási módszerek felhasználásával

TAKÁCS OLGA – VINCZE JÁNOS

ÖSSZEFOGLALÓ

Egy Blinder-Oaxaca típusú dekompozíciós nemi bérkülönbség elemzés eredményeit ismertetjük, amit magyar adatokon végeztünk el, és Véletlen Erdő regressziót használtunk. A nyers bérkülönbség a 2008-2016-os időszakban növekedett, de azt találtuk, hogy a bérstruktúra hatások stabilak voltak, és a nyers különbség növekedése az előzőleg negatív kompozíciós hatások eltűnésének tudható be. Vizuális elemzésünk érdekes nem-lineáris összefüggések meglétére utal, amelyek némelyikét jól tudjuk interpretálni. Egy Regressziós Fa elemzés azt sugallja, hogy bonyolult intearakciós mintákat rejtenek adataink. Identifikáljuk olyan szegmenseit a magyar munkapiacnak, amelyekben a legnagyobbak illetve a legkisebbek a bérstruktúra hatások. Elemzéseink alátámasztani látszanak azt az elméletet, amely szerint a nemi bérkülönbségek részben annak tulajdoníthatók, hogy a munkapiacok egy része monopszonisztikus, és eltérnek a férfiak és nők munkakínálati elaszticitásai.

JEL: J7 J3 C14

Kulcsszavak: nemi bérkülönbségek
Blinder-Oaxaca dekompozíció
Véletlen Erdő regresszió
Magyarország

The gender-dependent structure of wages in Hungary: results using machine learning techniques

Olga Takács
Corvinus University of Budapest

János Vincze
Corvinus University of Budapest
and
Centre for Economic and Regional Studies, Institute of Economics, (KRTK KTI)

November 2020

Abstract

This paper reports the results of a Blinder-Oaxaca style decomposition analysis on Hungarian matched employer-employee data to study the gender pay-gap. We carry out the decomposition by Random Forest regressions. The raw gap in our horizon (2008-2016) is increasing, but we find that the wage structure effects are rather stable, thus the rise in the gap is due to the disappearance of the formerly negative composition effects. Graphical analysis sheds light on interesting non-linear relationships; some of them can be readily interpreted by the previous literature. A Classification and Regression Tree analysis suggests that complicated interaction patterns exist in the data. We identify segments of the Hungarian labour market that are most and least exposed to gender-dependent wage determination. Our findings lend support to the idea that an important part of the gender wage gap is attributable to monopsonistic competition with gender-dependent supply elasticities.

Introduction

The existence of a gender wage gap, in favour of men, is an old and general issue. Many countries have legally endorsed the “Equal pay for equal work” principle [1]. The gap has recently narrowed worldwide, but it is still substantial. According to a study [2] women's average labour income in 2015 was 39 percent lower than men's in OECD countries, explaining why the narrowing of the gap is a major political target, for instance, in the European Union [3]. It is well understood that a mere average pay-gap is not necessarily caused by discrimination, which is a major underlying policy issue, and a huge literature has addressed the problem of identifying the discrimination component of the gap [4]. Our aim is not as ambitious as that, solely we would like to decompose and analyse the gap in a meaningful way, and, thereby, make inferences about the structure and operation of the labour market.

Perhaps the most frequently applied analytical tool for separating different components of the pay-gap has been the Blinder-Oaxaca (henceforward BO) decomposition [5], [6], [7]. As originally conceived, it breaks down the gap into an explained part, that displays the difference due to observed characteristics of workers, and an unexplained part, that is sometimes identified with the effect of discrimination [8], [9]. The discrimination interpretation presupposes a structural reference model of non-discriminatory wage determination. More and more researchers have lost faith in the possibility of this interpretation, giving several reasons from the lack of a well-established theory to the impossibility of observing relevant variables [10]. Even if one had a well-established empirical model for wage determination the empirical estimates would probably be inconsistent, due to selection bias, unobserved heterogeneity and errors-in-variables problems [11]. We agree with the conclusion expressed in [12], according to which the usual BO decomposition is not structural, and the unexplained gap cannot be construed as a measure of wage setting discrimination, in general. Still, this decomposition has proved very useful to establish important facts about the operation of labour markets, and the ensuing findings have been regarded as valuable indications for further research [12]. We will not use the “explained and unexplained gaps” terminology in this paper, rather we will refer to their equivalents as composition effects and wage structure (WS) effects, respectively [12].

The traditional BO decomposition is based on two regressions: the estimation of a reference model, which is, in most cases (see [4]), simply a model for determining men's wages, and a separate regression for women's wages. These regressions are best interpreted as approximations to the conditional expectation functions of wages. The idea that prompted our work is that it is not obvious that traditional regression techniques, such as OLS, are the best methods for this purpose. Statistical learning techniques (see [13]) have made inroads to econometrics, and it is worthwhile to experiment with them. We chose Random Forest (RF) regressions [14] as an alternative basis of the decomposition. RF is a tree-based statistical learning algorithm which has been applied in many disciplines [15]. Varian [13] proposed RF for econometricians by citing Howard and Bowles [16] who asserted that it has been one of the most successful general-purpose predictive algorithms. Wager and Athey [17] argued that RF regression is similar to other traditional non-parametric regression methods (e.g. k-nearest-neighbor algorithms), as it delivers some weighted average of “nearby” points as the prediction, but it has the advantage that both the weights and the proximities are determined in a data-driven way.

The usual BO decomposition results in a decomposition of the mean gap, which can be emulated by RF regressions. However, a more detailed analysis can be interesting, too. It is an acknowledged advantage of the OLS based decomposition that it leads naturally to a variable-wise decomposition of the wage structure and composition effects, though this is not as straightforward as it would seem to be [18]. As RF estimates do not yield parameters this route is not open to us. However, we can still estimate individual wage structure (IWS) effects (i.e. the hypothetical expected wage of a woman when “priced” as a man minus her expected wage when \priced" as a woman) with RF, and try to relate them to relevant covariates. Also, IWS effects can be used to identify segments of the labour market where these are extremely small or large. For this purpose we estimated Classification and Regression Tree (CART) models, see [19], with the estimated IWS effects as the target variable. CART is also a decision tree-based method that provides a sort of clustering in a supervised manner, providing a picture of the data in the space of observations rather than of variables, which is the usual point of view in econometrics. In a similar vein CARTs have been used for audience segmentation in public health research [20], to identify population subgroups whose members share common treatment effects.

Our analysis is conducted on Hungarian wage survey data for the years 2008-2016. By having data for nine consecutive years we can trace the time path of the composition and wage structure effects, and compare our findings with the literature that has dealt with similar problems, but applying the traditional methods [21]- [27].

In the next section we present our data and the statistical methodologies. The following section contains the results of the analysis, and the concluding section discusses the results in the light of the theoretical and empirical literature.

Data and methods

Data

Our data come from the Wage and Earnings Survey of the National Employment Office of Hungary, and were provided by the Databank of the Centre for Economic and Regional Studies. It is a matched employer-employee database that furnishes annual information (recorded in May). Each annual sample includes all firms with more than 50 employees and a randomly selected subset of firms with 5-50 employees. However, we dropped observations with firms having less than 20 employees as former research indicated that there is probably a large divergence between reported and actual wages in that size category [25], [28]. We used the logarithm of gross monthly earnings (called 'lnker' in the database), comprising the monthly base wage, overtime pay and other regular payments paid in May of each year, as the earnings variable. As this measure is inappropriate to compare full-time and part-time employees, we restricted our sample to employees working full-time. We left out the public sector, where wage setting is based on administrative rules. Table 1 shows the list of covariates used in our analysis.

Table 1. List of covariates (in parentheses the corresponding names in the database)

Name	***Unit***
Age (kor)	Years
Tenure (szolgho)	Months
Education (iskveg9_ordered)	1-9 categories, ordinal
Occupation (FEOR)	39 categories
Foreign control (kra_ordered)	1-4 ordinal
State control (ara_ordered)	1-4 ordinal
Firm size (letszam_bv1)	Number of employees
Settlement (ttip)	categorical, 1: Capital city, 2: Town, 3: Other

Region (kshreg)	7 categories, NUTS 2
Industry (ag1)	18 categories, NACE Rev. 2 - 1 digit
Collective labour agreement on firm level (kol)	0: no, 1: yes
Collective labour agreement on sectoral level (kag)	0: no, 1: yes
Collective agreement within several employers but not on sectoral level (ksz)	0: no, 1: yes

Notes: Tenure is the length of service with the current employer. Education refers to the highest completed level. The educational categories include: 1: Primary school 0-7 years, 2: Primary school 8 years,3: Vocational school, 4: Vocational training school, 5: Vocational high school, 6: Grammar school vocational education, 7: Technical institute, 8: Bachelor degree, 9: Master degree. Hungarian vocational, vocational training and vocational high schools combine general and vocational education, students learn general and professional lessons in different proportions. Vocational schools and vocational training schools don't provide secondary degree, while vocational high schools, grammar and technical schools do. Occupational code (FEOR) is the Hungarian variation of the 2-digit ISCO codes, see S1 Appendix. State (foreign) control consists of 4 categories: 1 is for 100%, 2 is for more than 50%, 3 is for less than 50% and 4 is for 0% of state (foreign) ownership. Industry categories are identical with the NACE Rev. 2 categories, for details see S1 Appendix. Regions (corresponding to NUTS 2 regions) are as follows: 1: Budapest and Pest county, 2: Central Transdanubia, 3: Western Transdanubia, 4: Southern Transdanubia, 5: Northern Hungary, 6: Northern Great Plain, 7: Southern Great Plain.

Source: Wage and Earnings Survey

The calculations were carried out using training and test samples from each year between 2008 and 2016. Each training sample contained 50 000 randomly selected observations, and the rest made up the test samples. Table 2 reports some basic statistics.

Table 2. Number of observations and the raw pay-gap in the dataset

Year	*Number of observations*	*Average wage gap*	*Female ratio in full dataset (%)*
2008	105509	0.1251	39.23
2009	95041	0.1137	40.23
2010	98174	0.1278	40.52
2011	98308	0.1379	40.26
2012	98654	0.1662	40.97
2013	101755	0.1413	38.31
2014	106986	0.1471	37.52
2015	131884	0.1513	38.51
2016	110003	0.16	39.5

Note: The raw gap is the difference between the average log wage of males and females.

Source: Wage and Earnings Survey

RF regressions

In RF regression one grows many suboptimal regression trees, and the RF prediction is calculated as an average of the individual trees' predictions. Each tree is grown from a bootstrap sample, and at each node only a random subset of explanatory variables are considered for a split. The main advantage of RF seems to be that the random and restricted manner of branch formation in individual trees achieves de-correlation among constituent trees, while unbiasedness is not jeopardized [29]. The full specification of the RF algorithm necessitates the setting of several parameters. After inspecting OOB (out-of-bag) prediction errors we decided that our forests contain 1000 trees each (see S2 Appendix). To control for the growth of individual trees we set the minimum node-size parameter at 5, and did not limit the maximum number of nodes. At every node the number of randomly selected variables was 5, out of 13 covariates. For the calculations we used the RandomForestSRC R package, which is based on [14].

BO decompositions

We calculated the decompositions with the male model as the reference. With the OLS methodology the following identity is valid:

$$\begin{aligned} av(y_M) - av(y_F) &= av(X_M) - av(X_F) \\ &= (av(X_M) - av(X_F))\beta_M + av(X_F)(\beta_M - \beta_F), \end{aligned} \tag{1}$$

where $av(y_M)$ and $av(y_F)$ are the average log earnings of groups labelled by M (male) and F (female), and $av(y_M) - av(y_F)$ is the difference of average male and female log wages (i.e. the raw gap). Here $av(X_M)$ and $av(X_F)$ denote the (vector) averages of the covariates in the subsamples β_M and β_F, while F are the respective OLS parameter vectors. In this equation the first term on the right-hand side is the composition effect, and the second measures the WS effect. When a constant is included in the OLS regressions the sample averages equal the average prediction, ae well known.

To carry out the BO-style analysis we estimated RF models on male and female subsets of the training samples. The RF prediction functions, P^M and P^F , were then applied to both the training and the test samples, divided into male and female subsets. For each man (woman) indexed by j (i), we got an estimated predicted wage $P^M(j)$ $(P^F(i))$. These estimates averaged over male (female) subsamples give $av(P^M(M))$ $(av(P^F(F)))$. The following identity holds:

$$av(y_M) - av(y_F) = av(P^M(M)) - av(P^F(F)) + bias, \tag{2}$$

where the arguments M and F refer to the identity of subsamples, and $av(P^M(M)) - av(P^F(F))$ is the predicted mean gender pay gap. The difference from the OLS based decomposition is in the non-zero bias term. However, this contrast disappears in the test samples where even OLS decompositions would contain non-zero bias. Thus, strictly speaking, we do not decompose the average differences but rather the average prediction differences. Denoting by $av(P^F(F))$ the average prediction for women when using the male prediction function, we obtain the following decomposition:

$$\begin{aligned} &av(P^M(M)) - av(P^F(F)) \\ &\qquad = [av(P^M(M)) - av(P^M(F))] + [av(P^M(F)) - av(P^F(F))], \end{aligned} \tag{3}$$

where the first term on the right-hand side is the composition effect, and the second is the WS effect. Clearly, if the wage-setting mechanisms, approximated by the male and female prediction functions, were the same and the predictions unbiased, then the first term would be equal to the raw gap. Otherwise, if the wage structure effects were non-zero, then we would think that the wage-setting mechanisms conditional on our predictors operate differently for the two genders. Notice that, besides the WS effect, the IWS (individual wage structure) effects, denoted by $e(k)$, can be estimated for each person k as the difference between predictions given by the male and female models:

$$e(k) = P^M(k) - P^F(k). \tag{4}$$

In the following we will refer to WS effects that are IWS effects averaged over some particular class of observations.

Analysis of the IWS effects

Aside from their role in the BO decomposition we could study the IWS effects independently, and relate them to specific variables. Looking for correlates we analysed graphically bivariate relationships between the estimated IWS effects and variables that had been identified in the literature as affecting the pay-gap. We will plot IWS effects against education levels, ownership, tenure, age, the femaleness of occupations and ISCO categories.

Finally, we wished to identify those sub-populations that exhibit the largest and the smallest IWS effects. We estimated CART [19] models with the estimated IWS effects as the target variable. Essentially, by growing a CART one subdivides the sample into homogenous groups, where homogeneity is defined via the dependent variable. CARTs balance two, countervailing, requirements: having a model that describes the data reasonably precisely, while providing a well-interpretable picture. Simple regression trees usually admit very clear interpretations, while more complex ones exhibit better fits. As our main goals was analysis, we did not go for the best possible predictive performance. Decent fits for CARTs are usually defined by the validation error curve. Researchers consider a model reliable if the complexity of the model is such that we enter the relatively flat part of this curve. Based on initial estimates we concluded that a complexity parameter, that controls implicitly the depth of the tree, of 0.001 would be a reasonable choice for all years (see S3 Appendix). We set the minimum number of observations in any leaf at 50, and used the default ten-fold cross validation option for calculating validation errors.

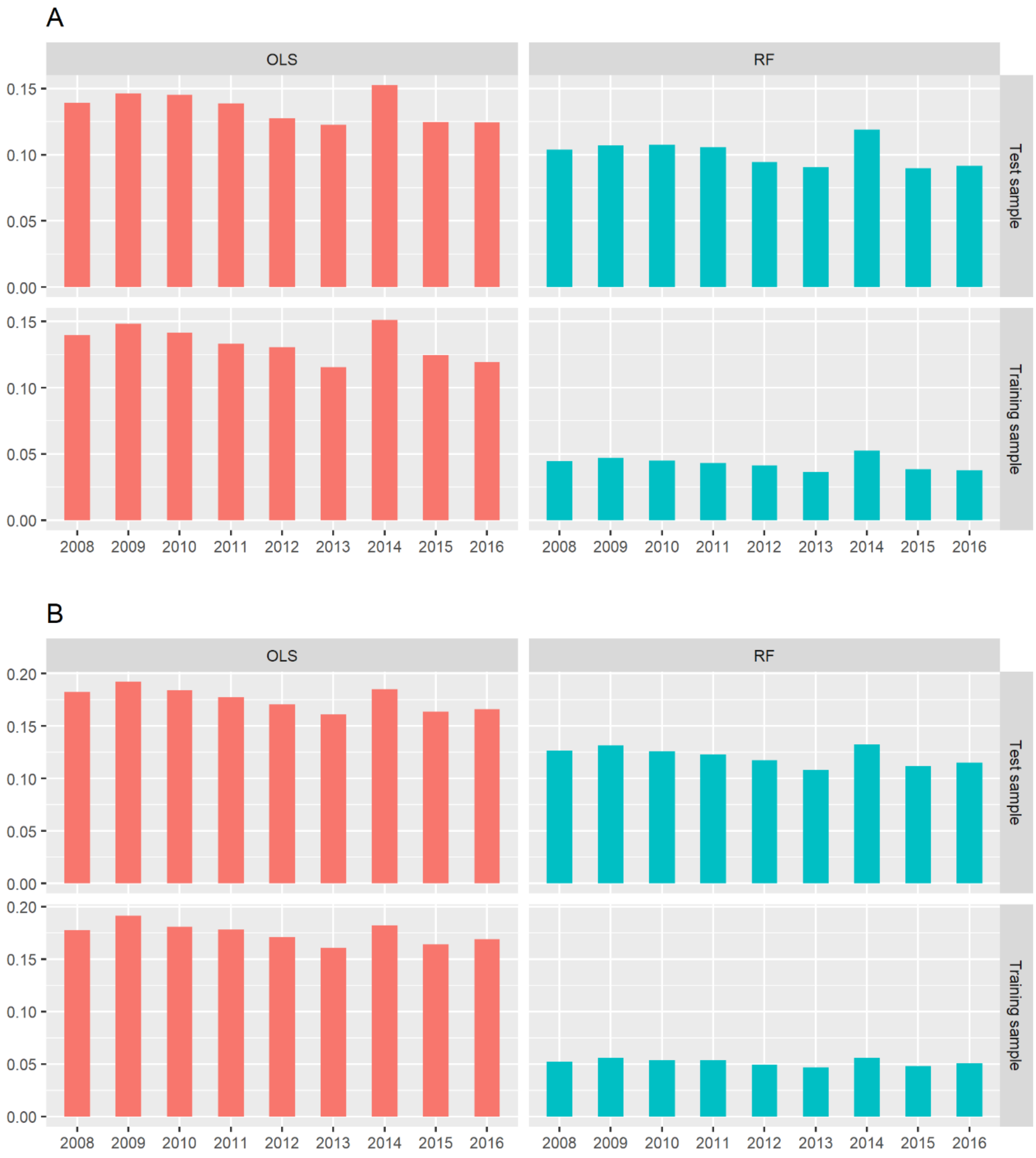
At first we experimented with the same set of variables that were used for the RF estimation, but it turned out that we could not obtain models with easily interpretable and reliable results. Therefore, we redefined variables in a way that all covariates had (not many) discrete values. Variables continuous in RF were redefined as ordinal, and we aggregated certain variables. To create the CARTs we used the rpart R-package that is based on [19].

Results

RF and OLS wage estimates: comparison of predictive performance

Needless to say it cannot be taken as an axiom that RF is better than OLS as a predictive device. Therefore, we have to document their relative performance for our particular dataset. The covariates described in Table 1 were used by the RF algorithms, while age-squared was also included in the OLS regressions as is customary in the literature [4]. We examined the predictive capability of the two methods by comparing MSEs for women and men, and on training and test samples, separately (see Fig 1).

Fig 1. MSE of RF and OLS estimates for training and test samples Panel A: women, Panel B: men



Source: Wage and Earnings Survey, own calculation

We find a much better fit by RF on the training datasets, which is not surprising since it is a non-parametric methodology. More importantly RF's better performance is observable on test data as well, for each year and for each gender. However, though MSEs do not perceivably increase from training to test data with the OLS estimates, there is a rise for the RF regressions. One can notice that MSEs are smaller in the female samples. All in all, we can conclude that RF seems to be at least as good, and probably better, data description tool than OLS, in our case. It must be noticed, however, the comparison between RF and OLS is not entirely fair in the sense that for a given set of explanatory variables many OLS (linear in parameters) models could, in principle, be specified, and we may not be clever enough to find the best specification. Thus we can claim only that RF seems to weakly outperform OLS with the usual specification.

The BO decompositions

The raw gaps in log points and their decompositions into composition effect and WS effect are presented in Table 3 and 4. (See also Fig 2 and Fig 3).

Table 3. BO decompositions of the log gender wage gap, training datasets

Year	***Raw wage gap***	***Composition effect***	***Wage structure effect***	***Bias***
2008	0.1203	-0.0361	0.1568	-0.0004
2009	0.1135	-0.0309	0.1448	-0.0004
2010	0.1358	-0.0216	0.1577	-0.0004
2011	0.1363	-0.0136	0.1504	-0.0005
2012	0.1692	0.0183	0.1514	-0.0005
2013	0.1466	-0.003	0.1501	-0.0006
2014	0.1431	0.0109	0.1329	-0.0008
2015	0.1500	0.0153	0.1349	-0.0003
2016	0.1716	0.0184	0.1537	-0.0005

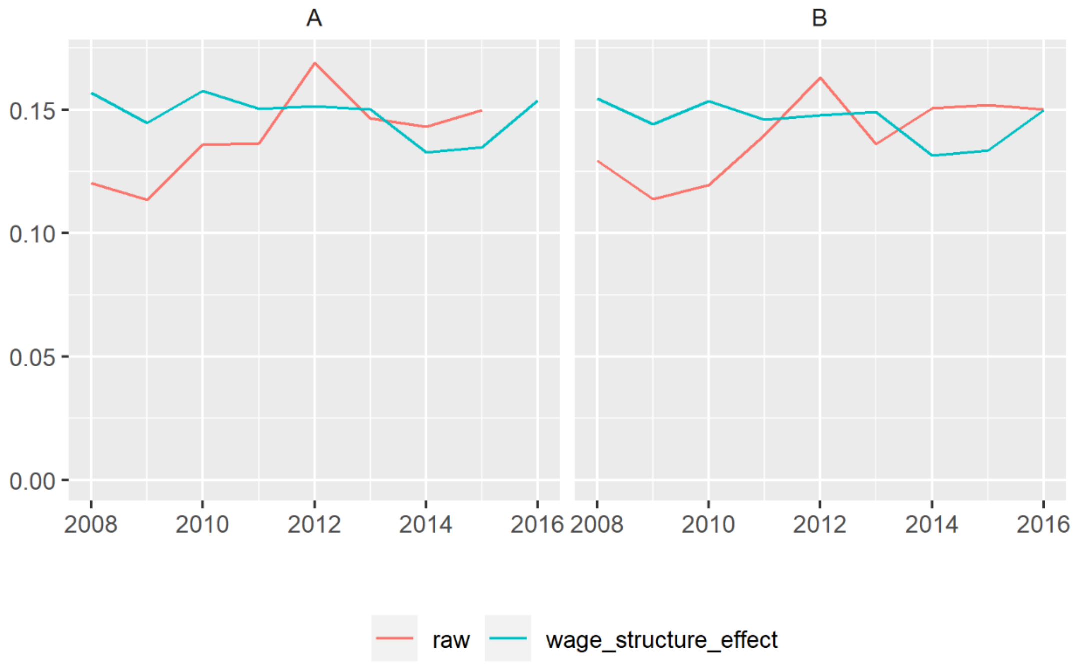
Source: Wage and Earnings Survey, own calculation

Table 4. BO decompositions of the log gender wage gap, test datasets

Year	***Raw wage gap***	***Composition effect***	***Wage structure effect***	***Bias***
2008	0.1295	-0.0257	0.1546	0.0006
2009	0.1139	-0.0318	0.1443	0.0014
2010	0.1196	-0.0283	0.1535	-0.0056
2011	0.1397	-0.014	0.1459	0.0078
2012	0.1631	0.0128	0.1478	0.0026
2013	0.1362	-0.0112	0.1491	-0.0017
2014	0.1507	0.0158	0.1315	0.0033
2015	0.1521	0.0118	0.1336	0.0066
2016	0.1503	0.0065	0.1500	-0.0062

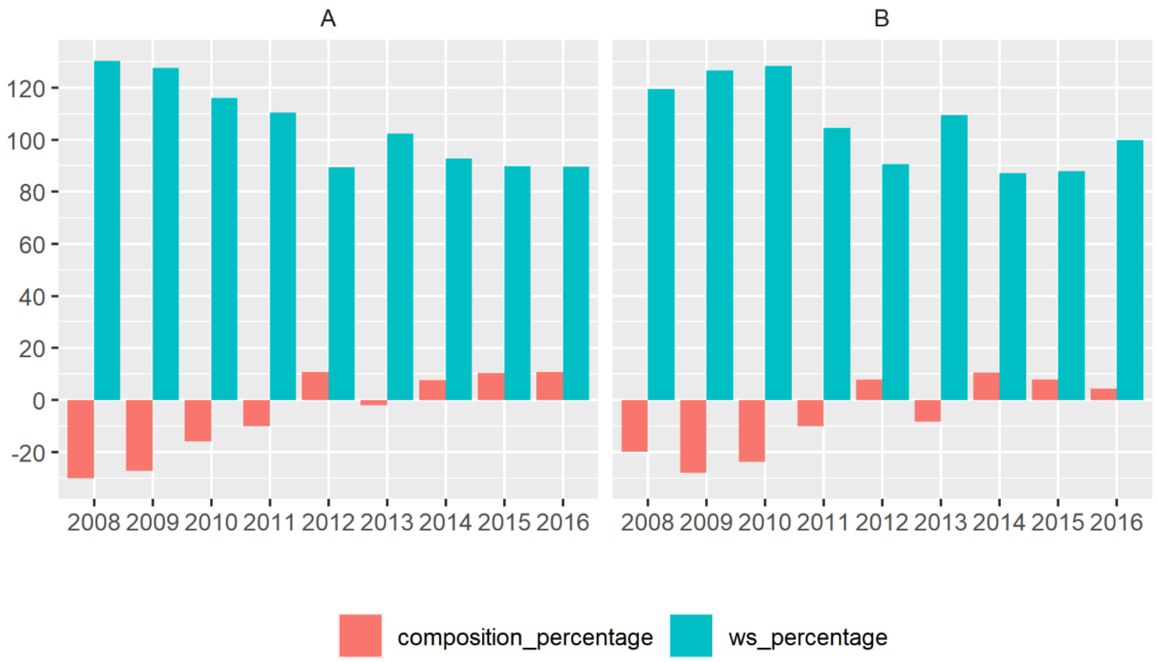
Source: Wage and Earnings Survey, own calculation

Fig 2. Raw gender wage gap and WS effects. Panel A: training data, panel B: test data



Source: Wage and Earnings Survey, own calculation

Fig 3. The BO decompositions (without biases) The effects are measured as percentages of the raw gap. Panel A: training data, panel B: test data



Source: Wage and Earnings Survey, own calculation

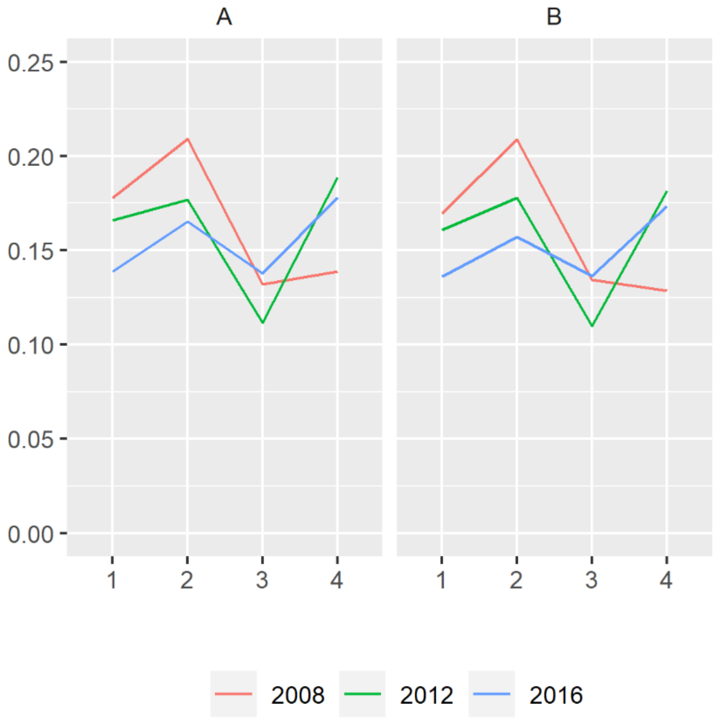
Over 2009-2016 the raw gap increased, though not monotonically. According to our calculations this upsurge can be attributed mainly to the change in the composition effect. The wage structure effects show a slight decline from 2011, whereas there is a pronounced shift in the composition effects. In the literature negative composition effects were found for Hungary and for several other countries [3], [31]. So it is not surprising that for 2008-2010 we obtained definitely negative composition effects. For 2011-2016 they are smaller in absolute value, and even positive in some years.

Correlates of the WS effects

In the following figures we plot conditional means of IWS effects with respect to a number of variables for three distinct years; 2008, 2012 and 2016. In all figures the covariates possess a natural ordering.

Fig 4 charts the relationship with the level of education, which is interestingly non-monotonic. The WS effect takes local maxima at secondary education without degree, i.e. it is larger in this category that either at the lowest educational level or at the next one (secondary education with degree). While for 2008 the maximum is global, for 2012 and 2016 tertiary education exhibits somewhat larger effects.

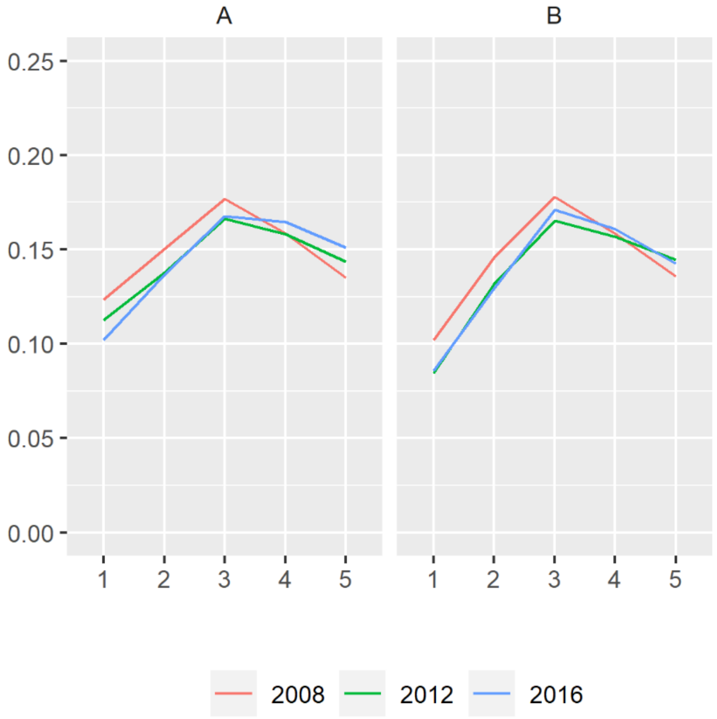
Fig 4. WS effects and education Education categories: 1: lower than secondary school, 2: secondary school without degree, 3: secondary school with degree, 4: tertiary degree



Source: Wage and Earnings Survey, own calculation

With respect to age Fig 5 displays a skewed inverted U, the effects increase until middle age, then decrease again for the oldest age groups, but the slope is smaller in the latter part of the graphs.

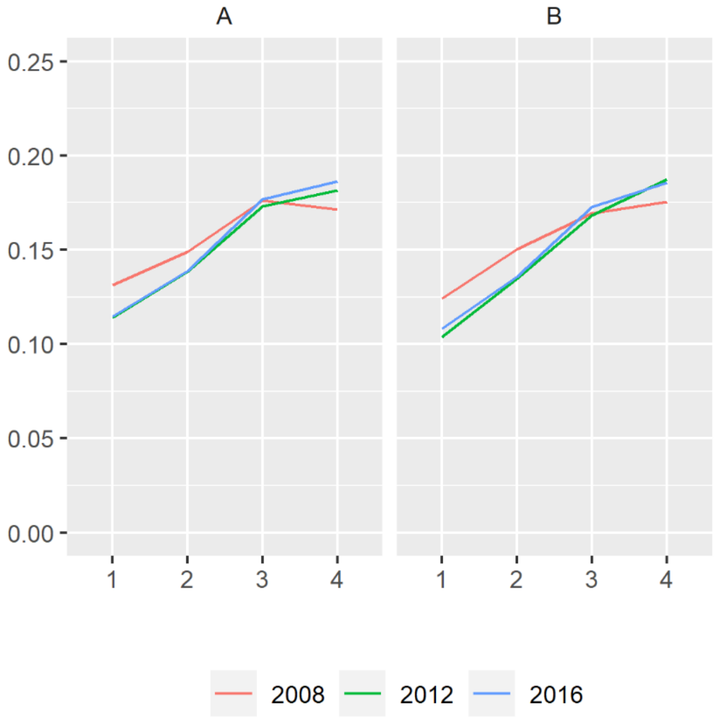
Fig 5. WS effects and age 1: Young (18-24 yrs), 2: Lower middle age (25-34 yrs), 3: Middle-age (35-44 yrs), 4: Upper middle age (45-54 yrs), 5: Older (55- yrs)



Source: Wage and Earnings Survey, own calculation

The relationship between tenure (time spent with the current employer) and WS effects seems to be (almost) unequivocally monotonically increasing (Fig 6), the effects are larger when women have worked longer in the same company. The only slight exception is 2008 with training data, where the fourth quartile assumes a somewhat smaller value than the third.

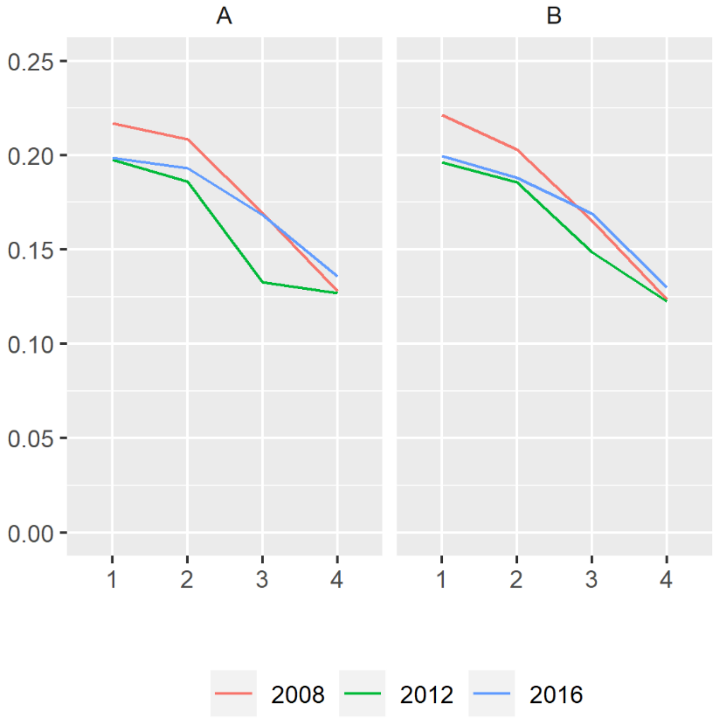
Fig 6. WS effects and tenure Tenure quartiles in the corresponding year



Source: Wage and Earnings Survey, own calculation

According to the testimony of Fig 7 majority foreign ownership is consistently associated with larger WS effects.

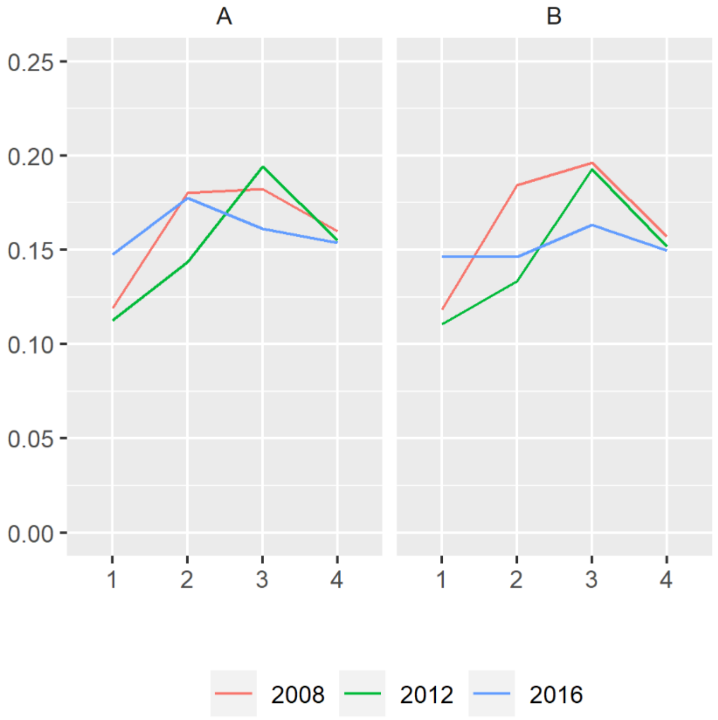
Fig 7. WS effects and foreign ownership Foreign ownership has 4 categories (see Table 1)



Source: Wage and Earnings Survey, own calculation

Fig. 7 provides information on the relationship with another type of ownership. It seems that full state property involves the smallest WS effects, while "no state property at all" is associated with somewhat larger effects than some state property (Fig 8). This figure is relatively chaotic, compared to the others; consistency in time is weak.

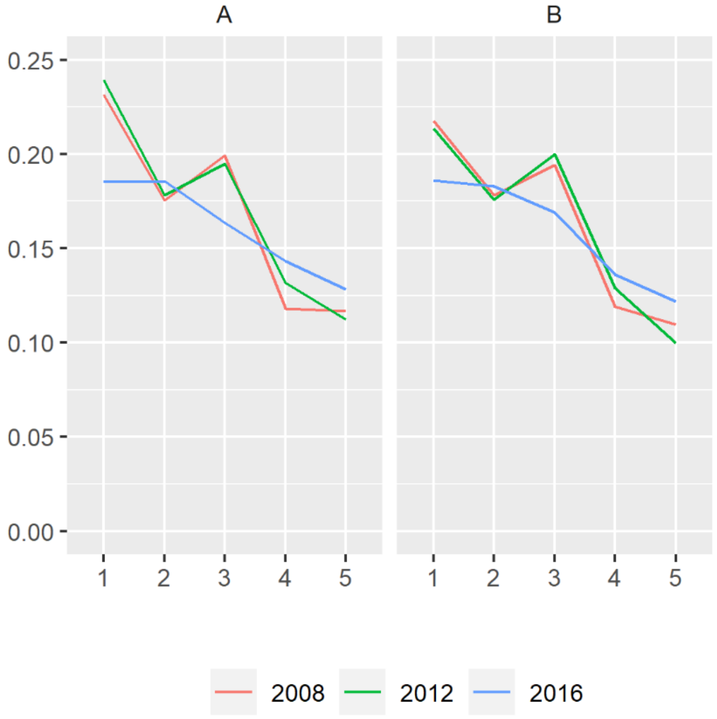
Fig 8. WS effects and state property State ownership has 4 categories (see Table 1)



Source: Wage and Earnings Survey, own calculation

In Fig 9 we relate the femaleness of an occupation to the WS effects. Apparently female dominated jobs are associated with smaller effects. It seems that women fare worst, by our measure, when they work in occupations where they constitute a minority, though a substantial minority (between 20 and 40 percent).

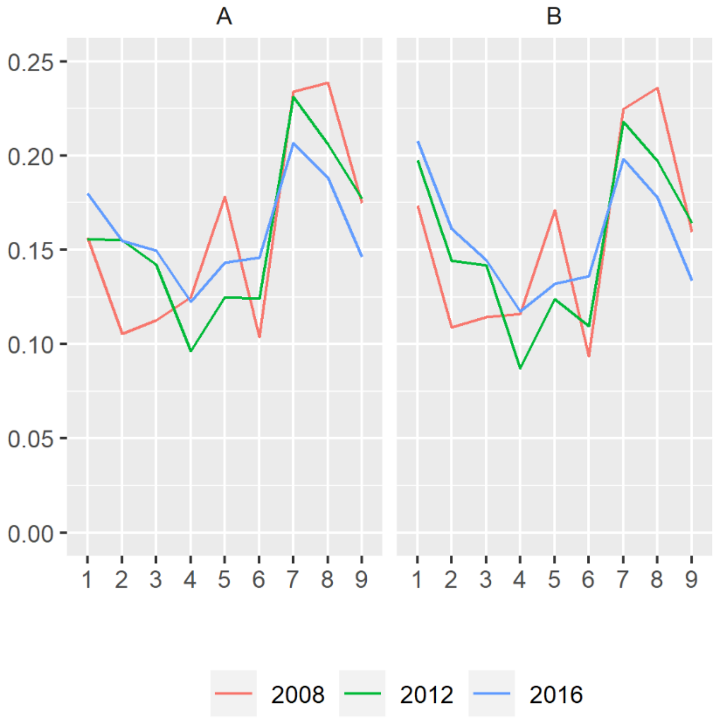
Fig 9. WS effects and the femaleness of occupations. Femaleness defined as share of women in two-digit ISCO occupational groups computed from the full Wages and Earnings Survey.



Source: Wage and Earnings Survey, own calculation

Finally, we looked into the association between the WS effects and one-digit ISCO categories (Fig 10). By and large, the first digits in the ISCO codes reflect decision making responsibility, lower numbers pertain to occupations with more managerial discretion. In 2008 and 2012 women in higher responsibility jobs tended to have smaller effects, but the difference faded by 2016. At one end "managers" (major group 1) had always larger effects than "professional occupations" (major group 2), whereas at the other end the largest effects showed up for "craft and related trades workers" (major group 7), rather than for those working in the simplest jobs (major group 9).

Fig 10. WS effects and occupations ISCO major groups 1 to 9. 1: Managers, 2: Professionals, 3: Technicians and associate professionals, 4: Clerical support workers, 5: Services and sales workers, 6: Skilled agricultural, forestry and fishery workers, 7: Crafts and related trades workers, 8: Plant and machine operators, and assemblers, 9: Elementary occupations



Source: Wage and Earnings Survey, own calculation

Extreme groups identified by CARTs

The above analysis of the estimated IWS effects is partial. To identify more complex relationships we ran CART models with the estimated IWS effects as the dependent variable and with covariates that are compressed versions of variables used in the RF regressions, with a view towards having tight and easily interpretable results. The categories are shown in Table 5.

Table 5. Covariates of IWS effects

Name	***Unit***
Age	5 categories
Tenure	4 quartiles
Education	4 categories
Foreign control	4 categories
State control	4 categories
Firm size	4 categories
Settlement	3 categories
Region	3 categories
Industry	4 categories
Collective agreement	0: no, 1: yes

Notes: Age is aggregated into five groups. 1: Young (18-24), 2: Lower middle age (25-34), 3: Middle-age (35-44), 4: Upper middle age (45-54), 5: Older (55-), and Tenure into quartiles. Education is categorized as 1: primary, 2: secondary school without degree, 3: secondary school with degree and 4: tertiary degree. Secondary school without degree includes vocational and vocational training schools. Secondary schools with degree are vocational high schools, grammar and technical institutions. Foreign and state control as well as settlement categories are the same as in Table 1. Firm size classes are as follows: Category 1: 20-49 employees, Category 2: 50-149 employees, Category 3: 150-499 employees, Categpry 4: 500- employees. Statistical regions are aggregated into 3 categories which corresponds to NUTS 1: Category 1 Central region (Budapest and Pest-county), Category 2 Transdanubian region (Central Transdanubia, Western Transdanubia and Southern Transdanubia), Category 3 Great Plain and North (Northern Hungary, Northern Great Plain and Southern Great Plain). Industry categorization is as follows. Category 1 (commerce): Wholesale and retail trade; repair of motor vehicles and motorcycles), Category 2: Manufacturing (highly tradable), Category 3 (somewhat tradables): Agriculture, forestry and fishing, Mining and quarrying, Electricity, gas, steam and air conditioning, Water supply, sewerage, waste management and remediation activities, Category 4 (non-tradables): Construction, Transportation and storage, Accommodation and food service activities, Information and communication, Financial and insurance activities, Real estate activities, Professional, scientific and technical activities, Administrative and support service activities, Education, Human health and social work activities, Arts, entertainment and recreation, Other service activities. Collective agreement is 1 if the employee has any kind of collective employment, 0 otherwise.

Source: Wage and Earnings Survey, own categorization

Unfortunately the CARTs produced too many leaves to admit a simple interpretation. Therefore we focus on extreme groups only. S4 Appendix presents the splitting attributes of the three groups (for each year) with the smallest WS effects (S-groups). These leaves represent the least under-priced types of women in our sample. We can see that, for many years, the WS effect is actually negative in some of these groups; in other words women belonging to them were priced, on average, higher than men with similar attributes. Inspecting Tables 1-9 in S4 Appendix with respect to sectoral distribution we find that manufacturing is almost never a splitting attribute. It can also be observed that the central region (where the capital city, Budapest, belongs to) appears rarely. Concerning firm size in most cases women working for firms with fewer employees show up usually in these groups. Little or no foreign ownership is many times a splitting attribute, and it seems that when education is a splitting variable the higher educational categories turn up, too. There is a certain time variation in the composition of the groups, the first five years seem to conform closely to the picture just described, there are differences in the years 2012 and 2013, then the pattern recurs apparently. Therefore, we can hypothesize that those women who are employed by the service sectors in the Central Region, by smaller and domestically owned companies, and who have upper-secondary or tertiary education might constitute a characteristic sub-population, the members of which are priced in the labour market roughly similarly to men, at least with respect to the attributes we observe in our data. More formally we defined a sub-population (S*) with these characteristics (see Table 6).

Table 6. S*-group characteristics

Variable	***Group***	***Code***
Region	central	1
Sector	non-manufactoring	not 2
Firm size	smaller (up to 150 employees	1,2
Education	higher secondary and tertiary	3,4
Foreign ownership	no foreign property	4

Source: Wage and Earnings Survey

In S4 Appendix (Tables 10-18) we exhibit also the splitting characteristics of the leaves with the three largest estimated WS effects (L-groups). These extreme groups contain the most under-priced women, where the degree of under-pricing is up to 30 percent sometimes. We can see that the variability over time between these groups is higher than that among the S-groups. Still certain features stand out, in particular, if we compare them with those at the other extreme. Concerning sectoral allegiance most women in L-groups work in manufacturing, in clear contrast with S-groups. Geographically the Transdanubian region, where a large part of the export oriented manufacturing industries have been settled, dominates. Also it seems that the firms that employ L-group women are frequently owned by foreigners. In sum, we may hypothesize that firms with majority foreign ownership in the export oriented manufacturing sector display the largest WS effects. It seems that firm size is not a consistently relevant variable. Concerning personal characteristics, in contrast to the S-groups, women with some, but not the highest, educational achievement appear most frequently. Formally we define our candidate L* sub-population in Table 7.

Table 7. L*-group characteristics

Variable	*Group*	*Code*
Region	Transdanubia	2
Sector	manufactoring	2
Education	primariy and lower secondary	1,2
Foreign ownership	foreign majority	1,2

Source: Wage and Earnings Survey

Though the S* and L* groups are unequivocally defined by the attributes given in Table 6 and Table 7 we should know all the relevant characteristics of these groups. Table 8 presents the attributes of the whole sample, while Table 9 and Table 10 portray the extreme groups.

Table 8. Characteristics of the total female population

Variable	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*
WS effect	0.1568	0.1448	0.1577	0.1504	0.1514	0.1501	0.1329	0.1349	0.1537
No.of obs.	19406	20184	20215	20119	20430	19076	18638	19364	19824
Age	39.81	40.1	40.43	40.43	40.55	40.48	40.41	40.53	41.02
Tenure	83.87	92.42	92.89	88.36	79.68	87.2	81.85	85.67	81.96
Education	5.09	5.26	5.23	5.28	5.27	5.3	5.35	5.38	5.43
Foreign	3.09	3.25	3.08	3.04	2.99	3	2.98	3.14	3.19
State	3.64	3.55	3.57	3.6	3.61	3.75	3.69	3.7	3.75
Firm size	1070.45	2993.74	2899.62	2769.99	2837.6	1393.2	2056.24	2484.61	1465.55
Collective	0.34	0.4	0.41	0.37	0.36	0.19	0.27	0.28	0.22

Source: Wage and Earnings Survey

Table 9. Characteristics of the observations in the S*-group

Variable	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*
WS effect	0.0497	0.0421	0.0388	0.0375	0.0515	0.0921	0.0633	0.0331	0.0775
No.of obs.	1886	2179	1778	1664	1448	1472	1622	1567	1879
Age	36.95	36.81	37.68	38.01	38.66	38.27	38.04	38.16	38.57
Tenure	53.67	53.54	60.53	60.25	64.1	61.69	56.39	55.79	57.39
Education	6.22	6.47	6.47	6.43	6.54	6.43	6.61	6.61	6.64
Foreign	4	4	4	4	4	4	4	4	4
State	3.7	3.81	3.82	3.78	3.82	3.83	3.83	3.88	3.89
Firm size	51.26	51.62	52.23	50.27	49.44	50.45	50.29	49.54	54.27
Collective	0.18	0.15	0.14	0.15	0.1	0.08	0.07	0.04	0.03

Source: Wage and Earnings Survey

Table 10. Characteristics of the observations in the L*-group

Variable	***2008***	***2009***	***2010***	***2011***	***2012***	***2013***	***2014***	***2015***	***2016***
WS effect	0.2704	0.2743	0.3193	0.2692	0.2374	0.2254	0.1805	0.2033	0.1913
No.of obs.	978	582	815	819	980	886	716	522	622
Age	40.88	41.74	41.52	42.07	41.25	42.06	42.65	41.95	43.52
Tenure	81.75	84.22	81.43	81.23	73.77	81.78	80.62	84.45	92.87
Education	3.04	3.14	3.18	3.11	3.2	3.3	3.06	3.17	3.16
Foreign	1.12	1.08	1.07	1.12	1.08	1.06	1.08	1.07	1.07
State	4	4	4	4	3.99	4	3.99	3.99	3.99
Firm size	1498.75	724.51	610.2	809.09	719.86	865.51	875.81	896.3	1090.8
Collective	0.33	0.35	0.3	0.31	0.44	0.32	0.22	0.2	0.18

Source: Wage and Earnings Survey

It can be seen that in S* the ratio of tertiary to upper-secondary educated women is larger than in the whole sample. Also age-group 2 (lower-middle age) has a higher share in S* than in the full sample. Concerning tenure, in each year the longest tenure quartile is underrepresented. It can be noticed that, maybe surprisingly, the ratio of workers with some collective agreement is lower than in the whole population. Regarding the group L* the middle-age and upper-middle-age groups (3 and 4) are somewhat overrepresented. Also, it can be observed, that the WS effects are much larger relative to the full sample average in the first years than in later years.

Discussion

To carry out the BO decomposition of the gender pay gap we estimated RF and OLS regression pairs on training samples: a male and a female model on the respective subsamples. We found that in each case (irrespective of time and gender) mean squared errors of RF on test data were smaller than mean squared errors of OLS on the training data. We feel vindicated that we proceeded to analyse the WS effects as measured by the RF regressions.

We found that though for the initial years a negative composition effect is estimated, it largely disappears from 2011, while the WS effect remains roughly constant, somewhat decreasing. It is likely that the initial negative composition effect was the long-term consequence of the pre-1990 (socialist economy) era, where long-term labour market decisions were made in a different economic and social environment that involved no substantial skill premium, therefore men were given fewer incentives to self-select into occupations requiring higher education [23]. There existed a concern in the literature (see [24] that the apparently higher educational achievement of women does not reflect real "productivity" advantages, since, among graduates, the subject of degree is also a relevant feature. As we we used also 2-digit ISCO codes, our findings are largely immune to this criticism. In addition, the radical transition in the 1990s brought a large drop in female labour market participation. As, naturally enough, low wage earners tended to exit, the average human capital characteristics of working women improved. It seems that by 2011 this effect of "initial conditions" evaporated.

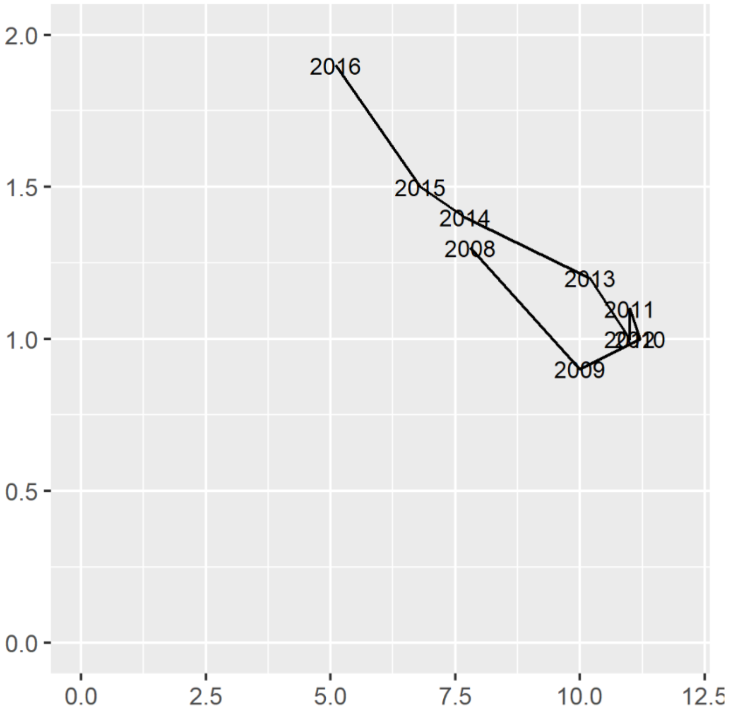
As the composition effect is due to selection bias, the WS effects can be regarded as the main concern for possible unequal treatment in paying labour, and it does not seem to have changed much after 2008. Therefore, the expectation that the development of the market economy would eventually reduce the WS effects (see [24] and [25]) has not materialized. Bivariate analyses of the individual IWS effects did not show substantial time variability, either. With respect to specific variables we found that educational achievement is nonlinearly associated with the IWS effects, and medium-skilled female workers' pay deficit is the largest. Concerning age our findings seem to accord well with the literature, as IWS effects increase first with age then slightly decrease, but still remain positive. The usual explanations rely on the lesser rate of human capital acquisition at the beginning of women' careers [32]. However, our finding that the WS effect increases with tenure does not square with previous findings in the literature [32]- [33]. It has been noted that multinational firms may price labour in a way that enhances the wage gap [34]. Our pertaining findings seem to corroborate this, as there seems to be a clear positive relationship between foreign ownership and the WS effects. State ownership does not appear to have any association with the WS effect, despite the tendency observed at the end of the 1990s that large and publicly owned firms exhibited relatively smaller gaps [24], [25].

The CART exercise demonstrated the existence of complex relationships. We found two characteristic sub-populations on either end of the IWS effects spectrum, which could not be identified by simply adding together the partial results. To understand the nature of these groups, we can invoke the monopsonistic labour markets theory [35]. With respect to the gender pay-gap this theory asserts that wage differentiation between sexes can be understood by reference to different labour supply elasticities of women and men. It relies on the idea that because of traditional family values earnings re less important determinants of the occupation choice for women, and women are also less mobile than men. These two features augment the market power of locally domineering large employers with respect to them. We can interpret our finding that an important under-priced sub-population consist of women working for foreign-owned manufacturing firms in the light of this theory. Territorially, the activity of these firms is concentrated in the Transdanubian region, and in smaller towns. Also, under-pricing affects most forcefully women working in less-skilled jobs, for whom salaries are anyway higher in these firms than they would be with alternative employers. Under-pricing also has a larger impact on those with a longer-attachment to the firm (longer tenure), as less mobility of women means that longer tenure does not represent transferable human capital enhancement for women compared to men.

In contrast, we observe the lowest levels of estimated WS effects in a sub-population consisting women who work in the service sector, at smaller firms, and in the Central Region that includes the capital city, Budapest. Here the demand side is not concentrated, thus monopsony power must be weaker. In addition, mobility across firms is easier even for women, and the importance of earning more may be greater for cultural reasons, especially for women with higher educational achievements. All of these together leave less room for offering lower wages to women than to men.

In general, if employers compete for workers more intensely the differences in individual elasticities tend to fade. If we look at the Hungarian Beveridge curve on Fig 11 we can see that after 2012 the Hungarian labour market became increasingly tight, which can be a reason why the WS effects may have been reduced somewhat in later years.

Fig 11. Hungarian Beveridge curve. Relationship between job vacancy (vertical axis) and unemployment rates (horizonzal axis) in Hungary between 2008 and 2016.



Source: Eurostat

Though these arguments cannot exclude more conventional causes, like unobserved human capital differences, or sheer gender discrimination, the huge differences between the IWS effects of these sub-populations make the unequal supply elasticities argument fairly convincing.

References

1. Blau FD, Kahn LM. Understanding international differences in the gender pay gap. Journal of Labor Economics 2003; 21(1) : 106-144.
2. OECD. Employment Outlook 2018. Paris: OECD publishing; 2018
3. Leythienne D, Ronkowski P. A decomposition of the unadjusted gender pay gap using Structure of Earnings Survey data. Luxembourg: Publications Office of the European Union; 2018
4. Weichselbaumer D, Winter{Ebmer R. A meta-analysis of the international gender wage gap. Journal of Economic Surveys. 2005; 19(3): 479-511.

5. Oaxaca, R. Male{female wage differentials in urban labor markets. International Economic Review. 1973; 14: 693-709.

6. Blinder, A. S. Wage discimination: Reduced form and structural estimates. Journal of Human Resources. 1973; 8: 436-455.

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

8. Altonji JG, Blank RM. Race and Gender in the Labor Market. In: Ashenfelter, O.-Card, D., editors. Handbook of Labor Economics. Amsterdam: Elsevier; 1999. pp. 3144-3259

9. Darity WA, Mason PL. Evidence on discrimination in employment: Codes of color, codes of gender. Journal of Economic Perspectives, 1998; 12(2): 63-90. doi: http://dx.doi.org/10.1257/jep.12.2.63

10. Weichselbaumer D, Winter{Ebmer R. Rhetoric in economic research: The case of gender wage differentials. Industrial Relations: A Journal of Economy and Society. 2006; 45(3): 416-436. doi: http://dx.doi.org/10.1111/j.1468-232X.2006.00431.x

11. Kunze, A. Gender wage gap studies: consistency and decomposition. Empirical Economics, 2008, 35(1) 63-76.

12. Fortin N. Lemieux T, Firpo S. Decomposition methods in economics. In Handbook of Labor Economics 4: 1-102. Elsevier; 2011.

13. Varian HR, Big data: New tricks for econometrics. Journal of Economic Perspectives. 2014; 28(2): 3-28 doi: http://dx.doi.org/10.1257/jep.28.2.3

14. Breiman L. Random forests. Machine Learning, 2001; 45(1): 5-32. doi: https://doi.org/10.1023/A:1010933404324

15. Cutler A, Cutler DR, Stevens JR. Random forests. InEnsemble machine learning 2012 (pp. 157-175). Springer, Boston, MA.

16. Howard J, Bowles M. The two most important algorithms in predictive modelling today. In Strata Conference presentation, 2012; 28(2)

17. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association. 2018; 113(523): 1228-1242

18. Gardeazabal J, Arantza U. More on identification in detailed wage decompositions. Review of Economics and Statistics. 2004; 86(4): 1034-1036

19. Breiman L, Friedman J, Stone C J, Olshen RA. Classification and regression trees. USA: CRC Press; 1984

20. Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., and Rakowski, W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Annals of Behavioral Medicine 2003;26(3), 172-181.

21. Lovász A. Jobbak a nők esélyei a közszférában? A nők és férfiak bérei közötti Különbség és a foglalkozási szegregáció vizsgálata a köz- és a magánszférában. (Are the chances of women better in the public sector? Gender wage differences and segregation in the public and private sectors) 2013; 40(7-8): 814-836. Hungarian.

22. Brainerd, E Women in transition: Changes in gender wage differentials in Eastern Europe and the former Soviet Union. ILR Review, 2000; 54(1), 138-162. http://dx.doi.org/10.2307/2696036

23. Newell A, Reilly B. The gender pay gap in the transition from communism: some empirical evidence. Economic Systems. 2001; 25(4): 287-304

24. Jolliffe D, Campos NF. Does market liberalisation reduce gender discrimination? Econometric evidence from Hungary, 1986{1998. Labour Economics. 2005;12(1):

1-22.

25. Lovász A. Competition and the Gender Wage Gap: New Evidence from Linked Employer-Employee Data in Hungary, 1986-2003. Budapest Working Papers On The Labour Market BWP. 2008 July

26. Cukrowska E, Lovász A. Are children driving the gender wage gap? Comparative evidence from Poland and Hungary. Budapest Working Papers On The Labour Market BWP. 2014; Issue 4.

27. Csillag M. "Female work" and the gender wage gap from late socialism to today. In Galasi P. and Kézdi G.(editors): The hungarian labour market review and analysis, Budapest, Institute of Economics, 2007

28. Elek P, Scharle Á, Szabó B, Szabó PA. A bérekhez kapcsolódó adóeltitkolás Magyarországon. (Tax evasion and wages in Hungary) Közpénzügyi Füzetek. 2009; 23. Hungarian.

29. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. New York: Springer-Verlag; 2009

30. Olivetti C, Petrongolo B. Unequal pay or unequal employment? A cross-country analysis of gender gaps. Journal of Labor Economics. 2008; 26(4): 621-654. doi: http://dx.doi.org/10.1086/589458

31. Machin S, Puhani P.A. Subject of degree and the gender wage differential: Evidence from the UK and Germany. Economics Letters 2003; 79.3: 393-400.

32. Card D, Cardoso AR, Kline P. Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. The Quarterly Journal of Economics. 2015; 131(2): 633-686.

33. Kunze A. The evolution of the gender wage gap. Labour Economics. 2005; 12(1): 73-97.

34. Vahter P, Masso J. The contribution of multinationals to wage inequality: foreign ownership and the gender pay gap. Review of World Economics; 2019; 155(1): 105-148.

35. Manning, A. Monopsony in motion: Imperfect competition in labor markets. Princeton; Oxford: Princeton University Press; 2003. Chapter 7

Appendix

S1 Tables Occupational codes and industry categories

Table 1. Occupational codes

ISCO code	*Name*
11	Chief executives, senior officials and legislators
12	Managing directors and chief executives of business organisations and budgetary…
13	Production and specialized services managers
14	Heads of units assisting business activities
21	Technical, information technology and science related professionals
22	Health professionals
23	Social services professionals
24	Educators, teachers
25	Business type professionals
26	Legal and social sciences professionals
27	Culture, sports, arts and religion professionals
29	Other highly qualified executives
31	Technicians and other related technical professionals
32	Supervisors
33	Health professionals
34	Educational assistants
35	Social health care and labour market services professionals
36	Business related services administrators, administrators of authorities, agents
37	Arts, cultural, sports and religious professionals
39	Other administrators
41	Office clerks
42	Customer services occupations
51	Commercial and catering occupations
52	Service workers
61	Agricultural occupations
62	Forestry, game-farming and fisheries occupations
71	Food processing workers
72	Light industry occupations
73	Metal and electrical industry occupations
74	Handicraft workers
75	Building industry occupations
79	Other industry and construction industry occupations
81	Manufacturing machine operators
82	Assemblers
83	Stationary machine operators
84	Drivers and mobile machinery operators
91	Cleaners and related simple occupations
92	Simple service, transport and similar occupations
93	Simple industry, construction industry, agricultural occupations

Source: Central Statistical Office

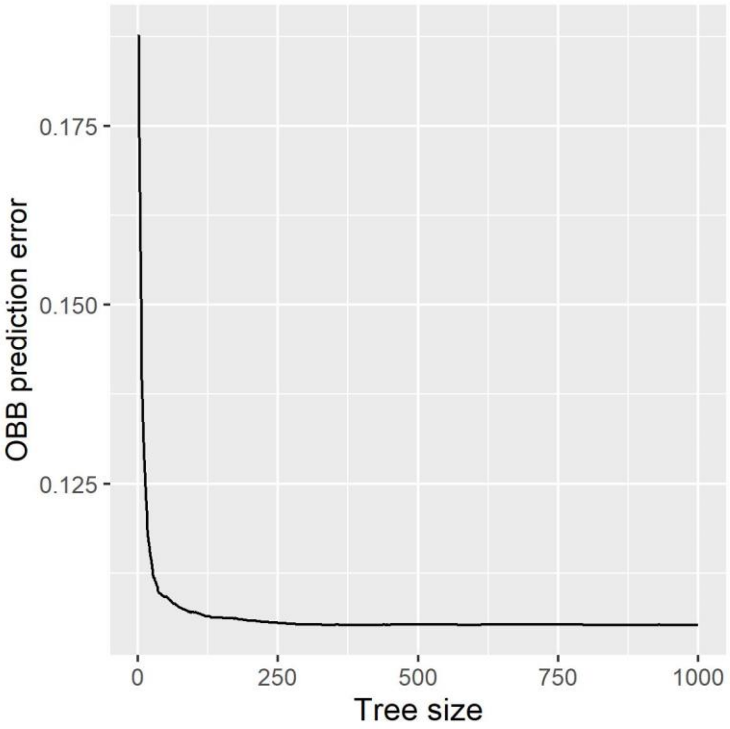
Table 2. Industry codes

Industry	***Name***
A	Agriculture, forestry and fishing
B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam and air conditioning supply
E	Water supply; sewerage, waste management; remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	Transportation and storage
I	Accommodation and food service activities
J	Information and communication
K	Financial and insurance activities
L	Real estate activities
M	Professional, scientific and technical activities
N	Administrative and support service activities
O	Public administration and defence; compulsory social security
P	Education
Q	Human health and social work activities
R	Arts, entertainment and recreation
S	Other service activities
T	Activities of households as employers
U	Activities of extraterritorial organisations and bodies

Source: Central Statistical Office

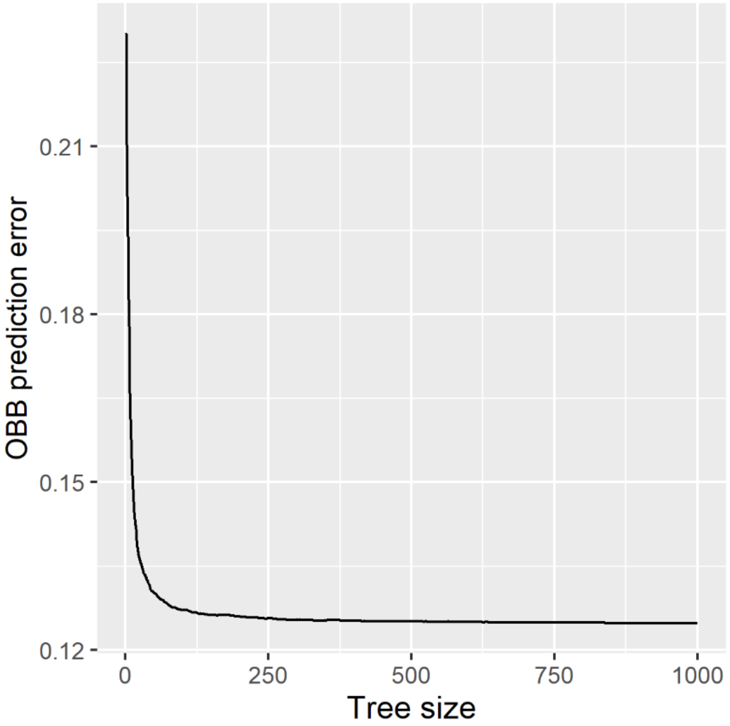
S2 Appendix OOB prediction errors for RF regressions

Fig 1. OOB prediction errors of RF for women in 2008



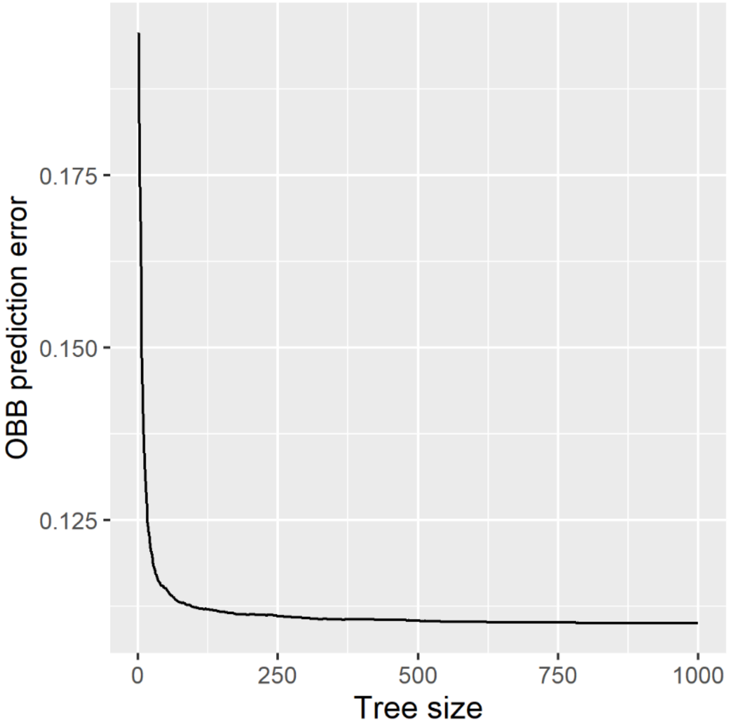
Source: Wage and Earnings Survey, own calculations

Fig 2. OOB prediction errors of RF for men in 2008



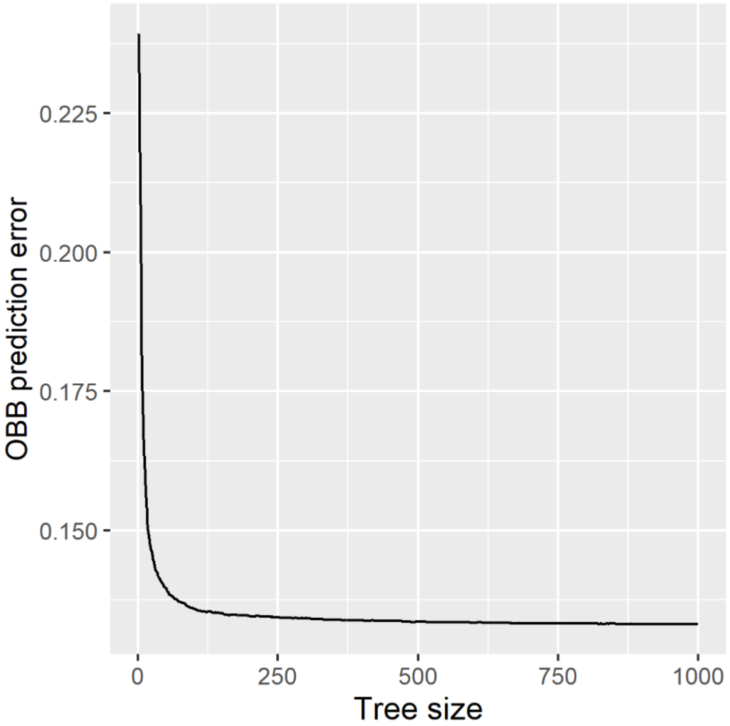
Source: Wage and Earnings Survey, own calculations

Fig 3. OOB prediction errors of RF for women in 2009



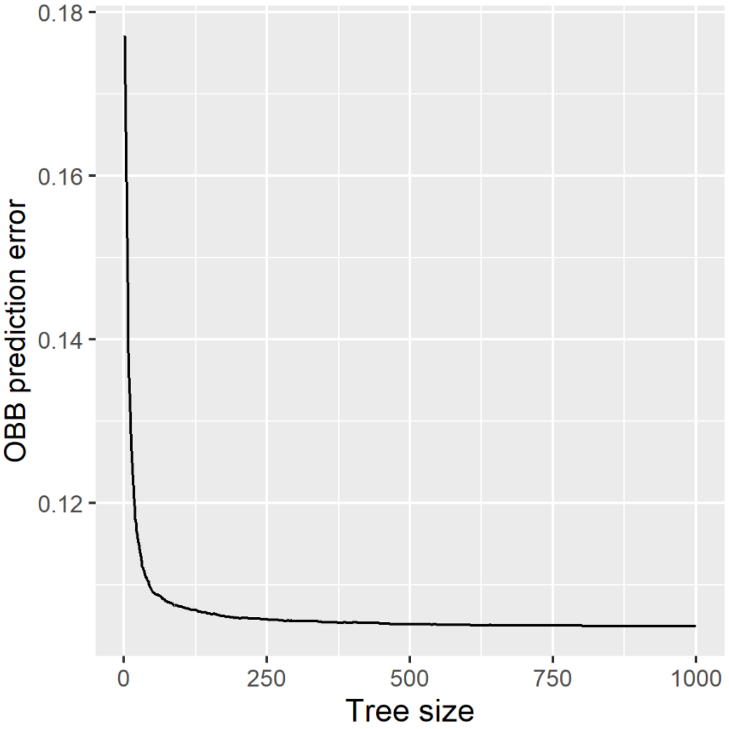
Source: Wage and Earnings Survey, own calculations

Fig 4. OOB prediction errors of RF for men in 2009



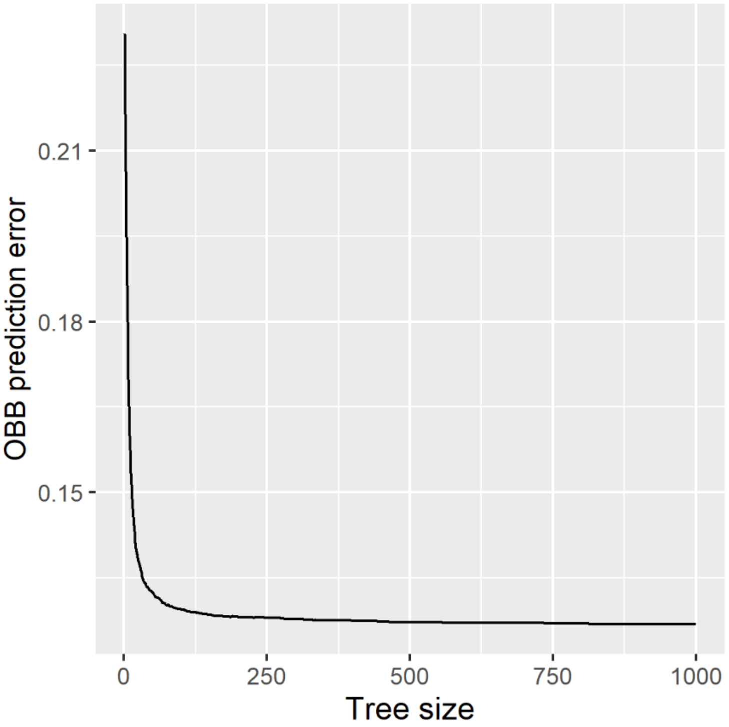
Source: Wage and Earnings Survey, own calculations

Fig 5. OOB prediction errors of RF for women in 2010



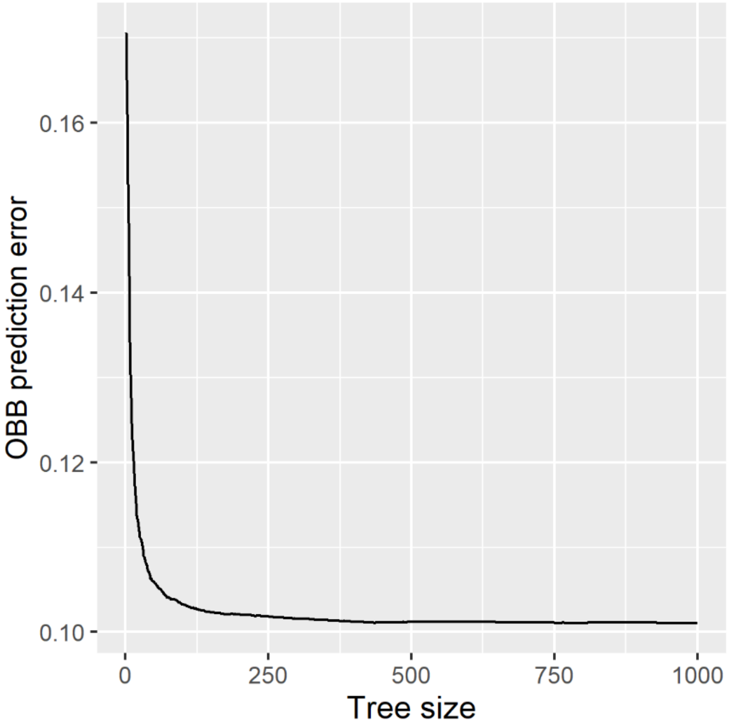
Source: Wage and Earnings Survey, own calculations

Fig 6. OOB prediction errors of RF for men in 2010



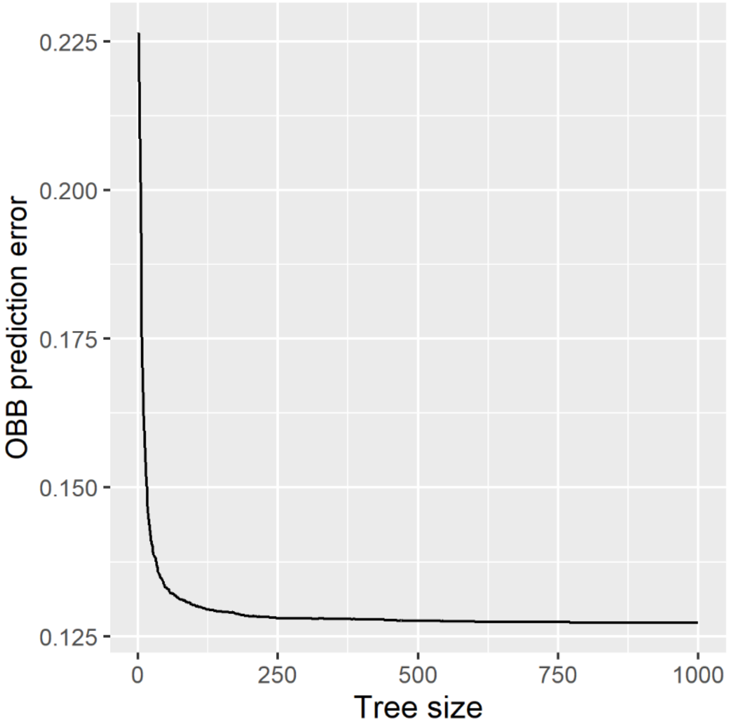
Source: Wage and Earnings Survey, own calculations

Fig 7. OOB prediction errors of RF for women in 2011



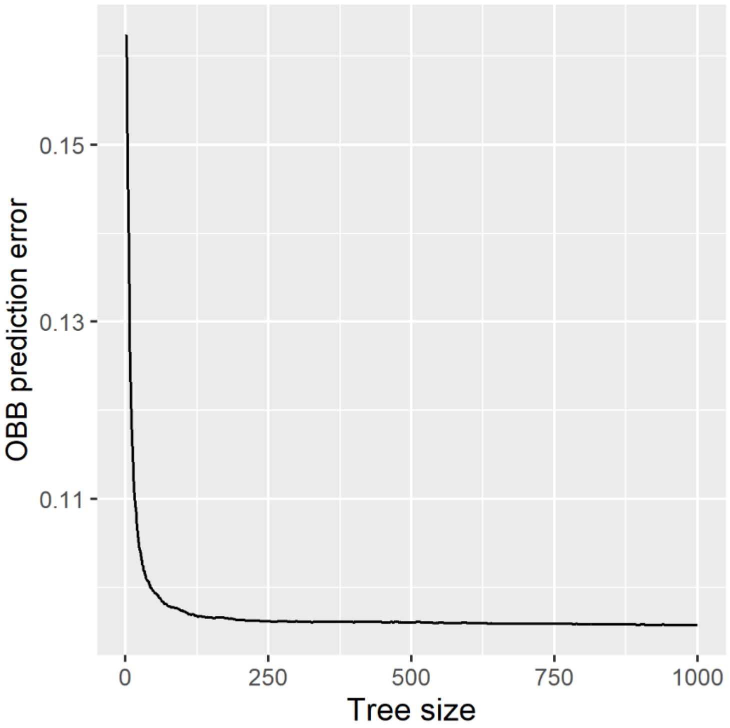
Source: Wage and Earnings Survey, own calculations

Fig 8. OOB prediction errors of RF for men in 2011



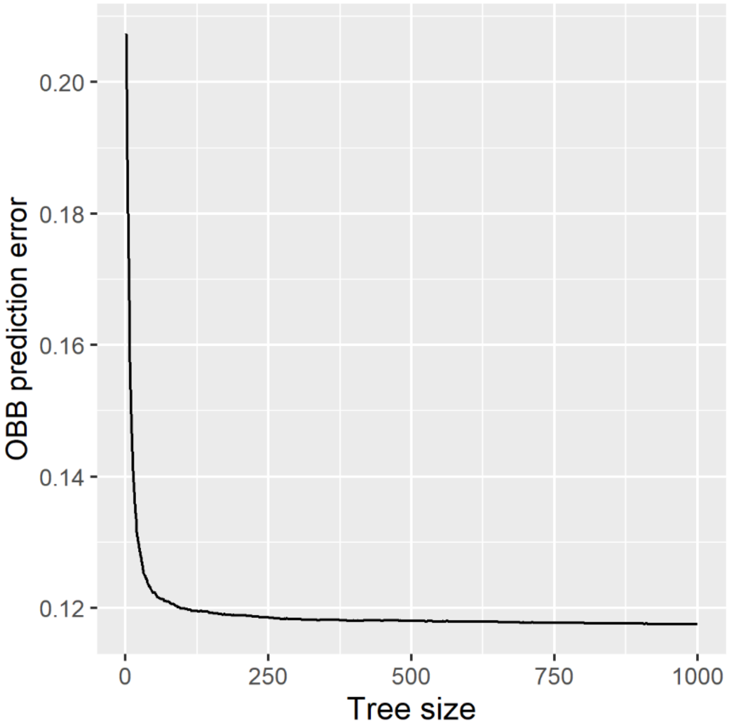
Source: Wage and Earnings Survey, own calculations

Fig 9. OOB prediction errors of RF for women in 2012



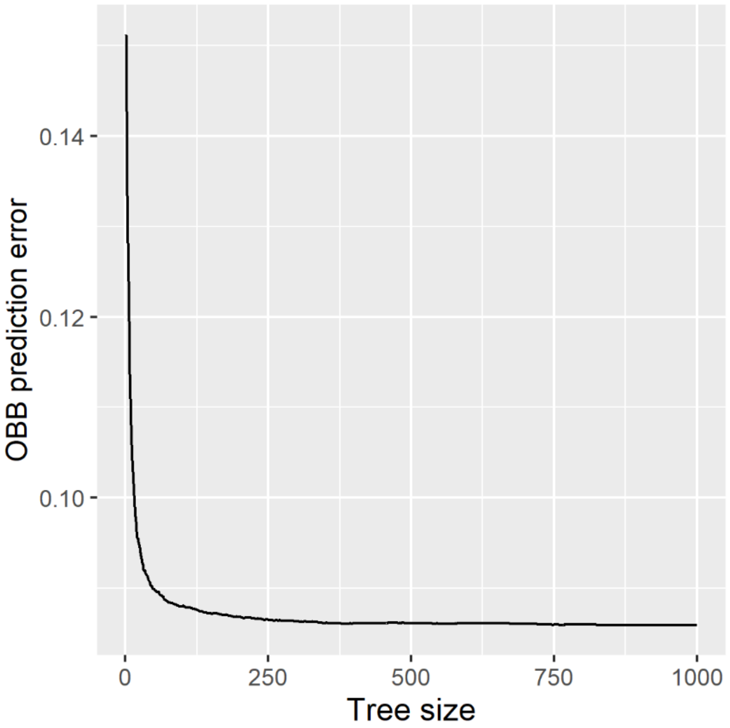
Source: Wage and Earnings Survey, own calculations

Fig 10. OOB prediction errors of RF for men in 2012



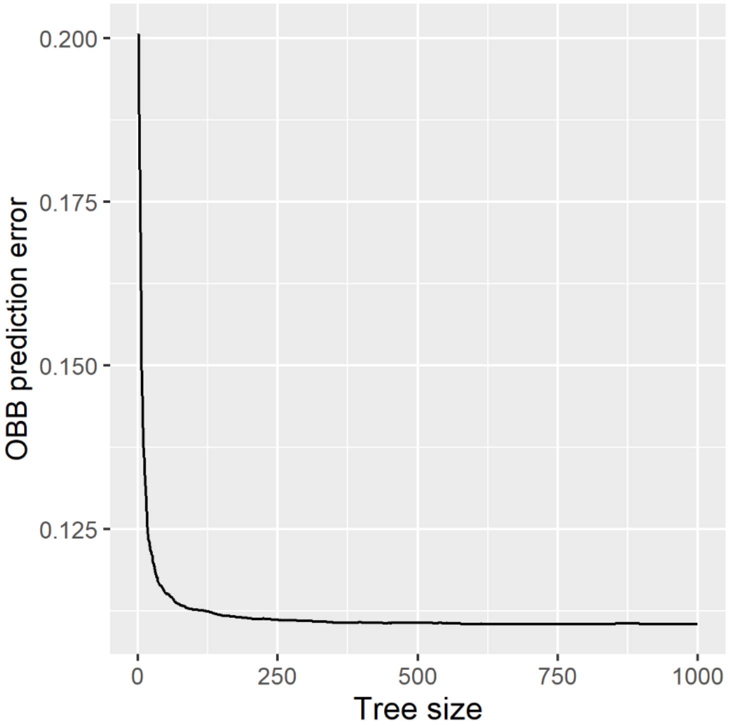
Source: Wage and Earnings Survey, own calculations

Fig 11. OOB prediction errors of RF for women in 2013



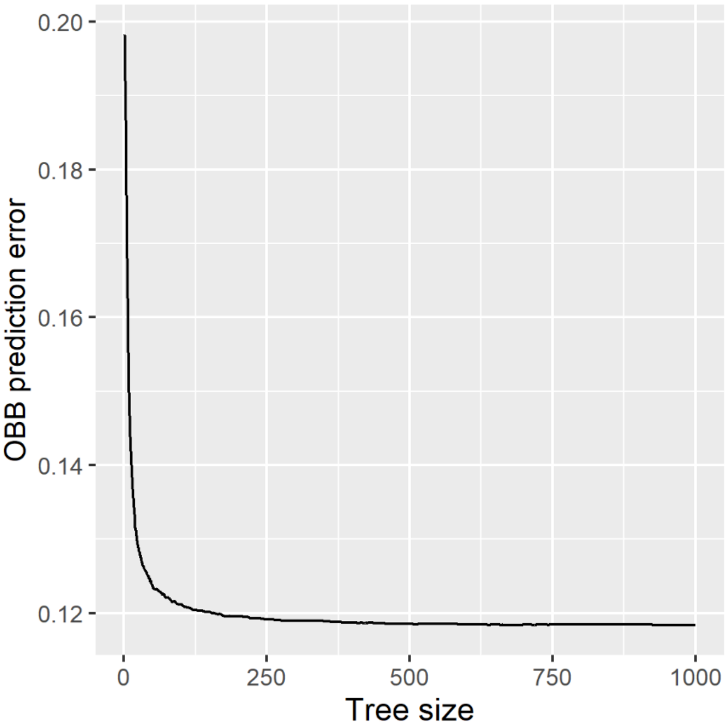
Source: Wage and Earnings Survey, own calculations

Fig 12. OOB prediction errors of RF for men in 2013



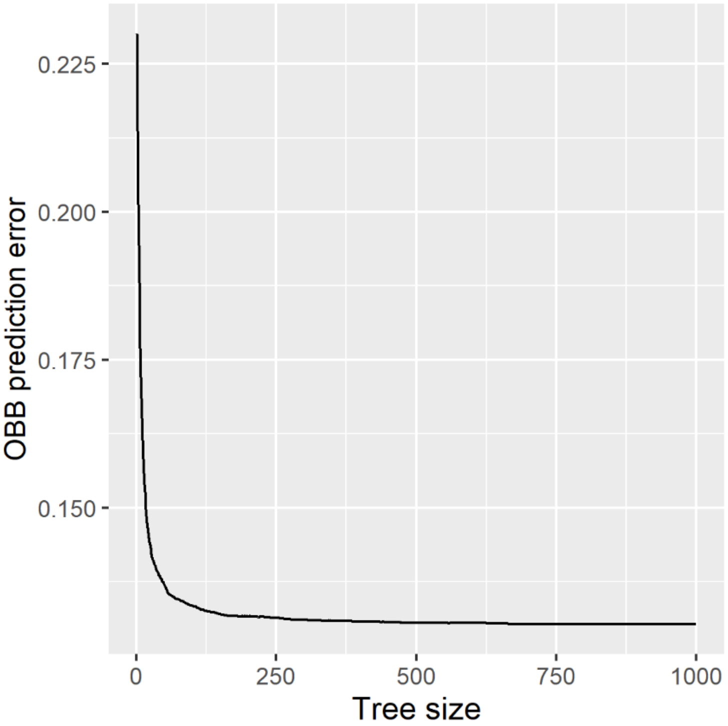
Source: Wage and Earnings Survey, own calculations

Fig 13. OOB prediction errors of RF for women in 2014



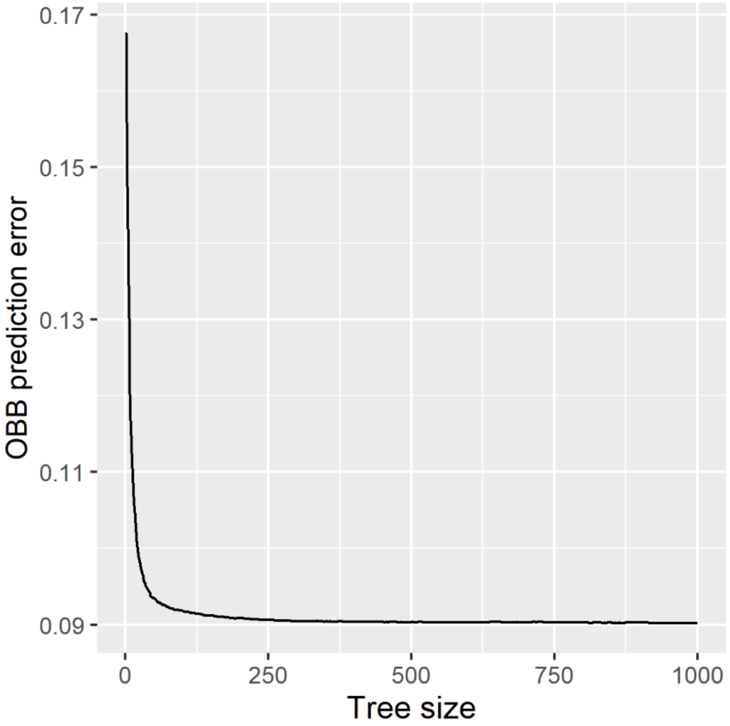
Source: Wage and Earnings Survey, own calculations

Fig 14. OOB prediction errors of RF for men in 2014



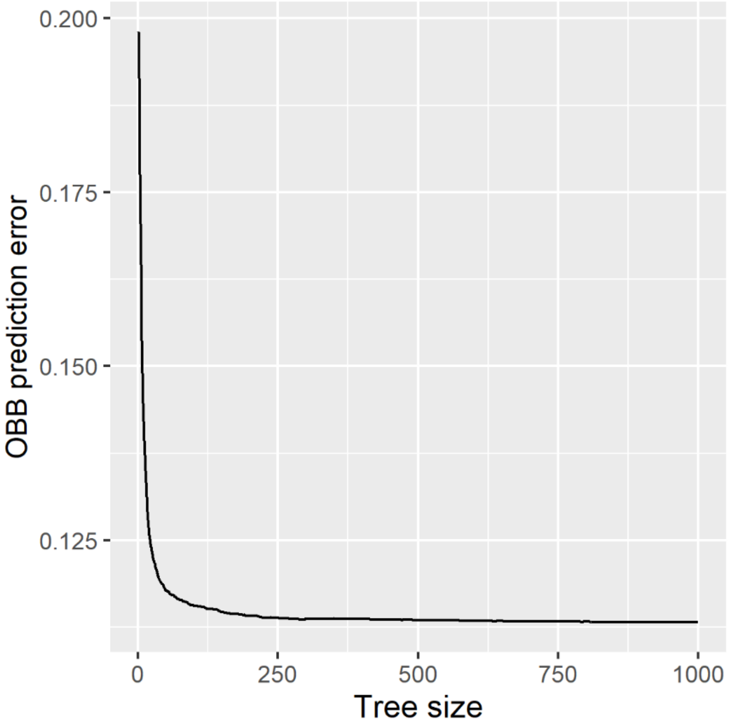
Source: Wage and Earnings Survey, own calculations

Fig 15. OOB prediction errors of RF for women in 2015



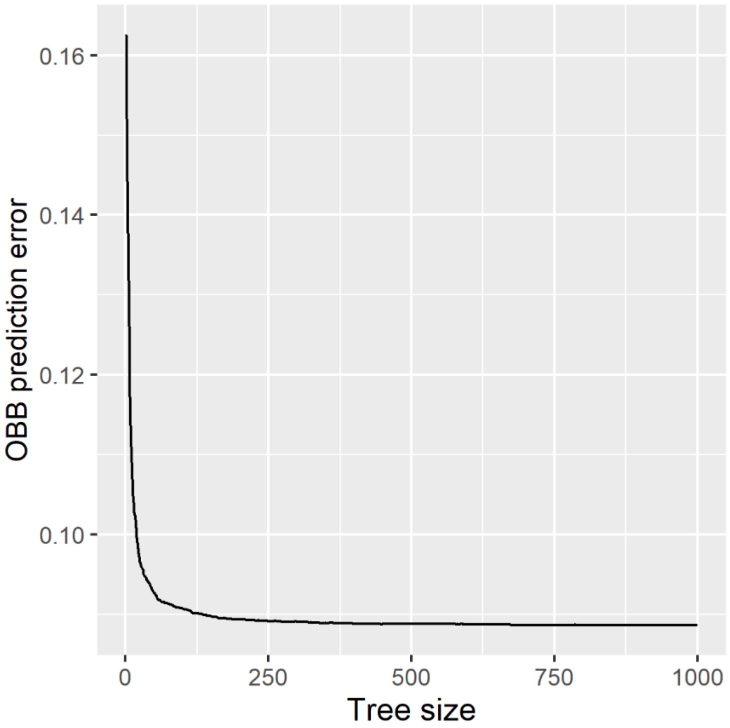
Source: Wage and Earnings Survey, own calculations

Fig 16. OOB prediction errors of RF for men in 2015



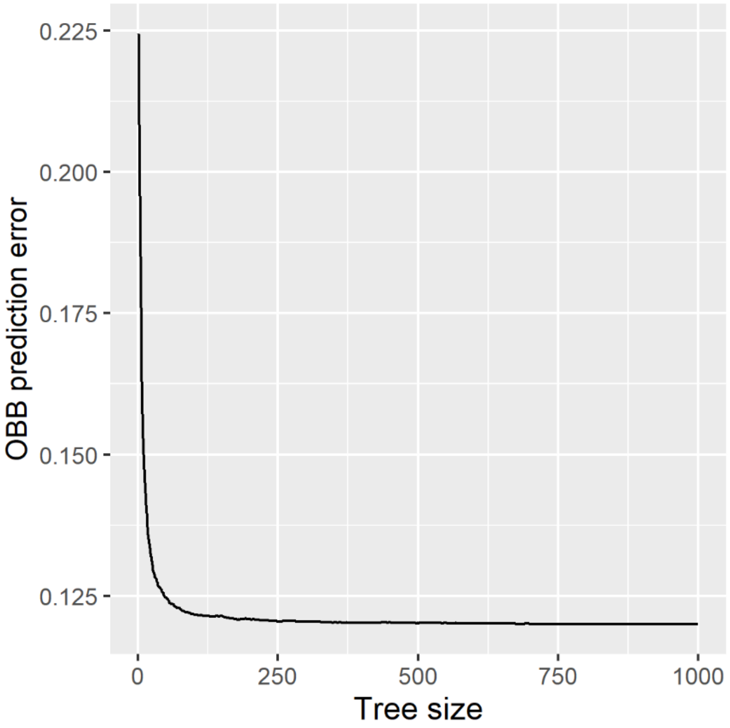
Source: Wage and Earnings Survey, own calculations

Fig 17. OOB prediction errors of RF for women in 2016



Source: Wage and Earnings Survey, own calculations

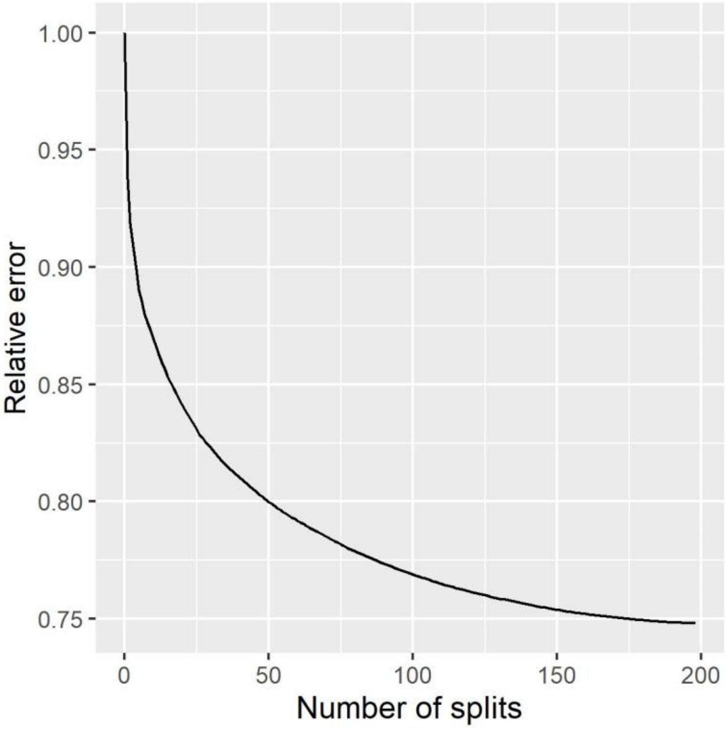
Fig 18. OOB prediction errors of RF for men in 2016



Source: Wage and Earnings Survey, own calculations

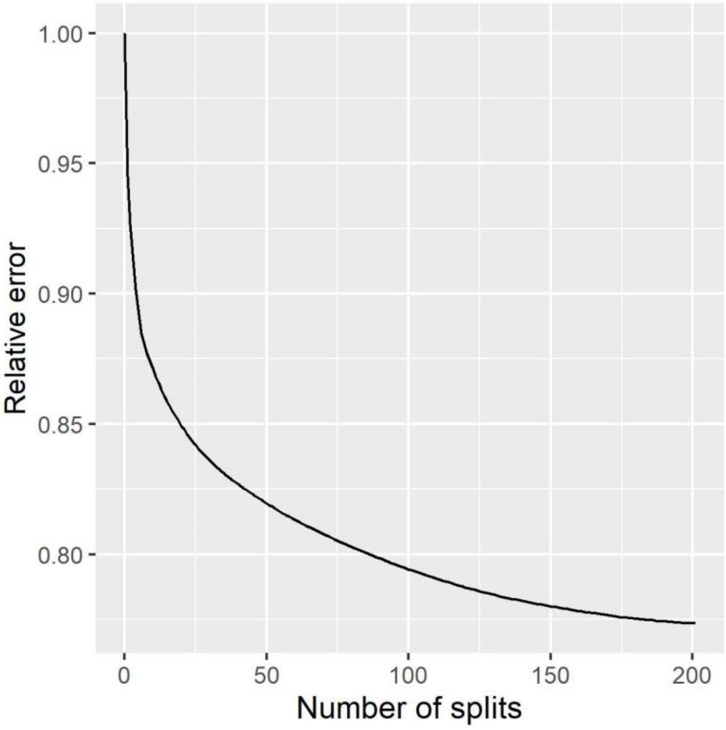
S3 Appendix Validation error curves for CARTs

Fig 1. Validation error curve in 2008



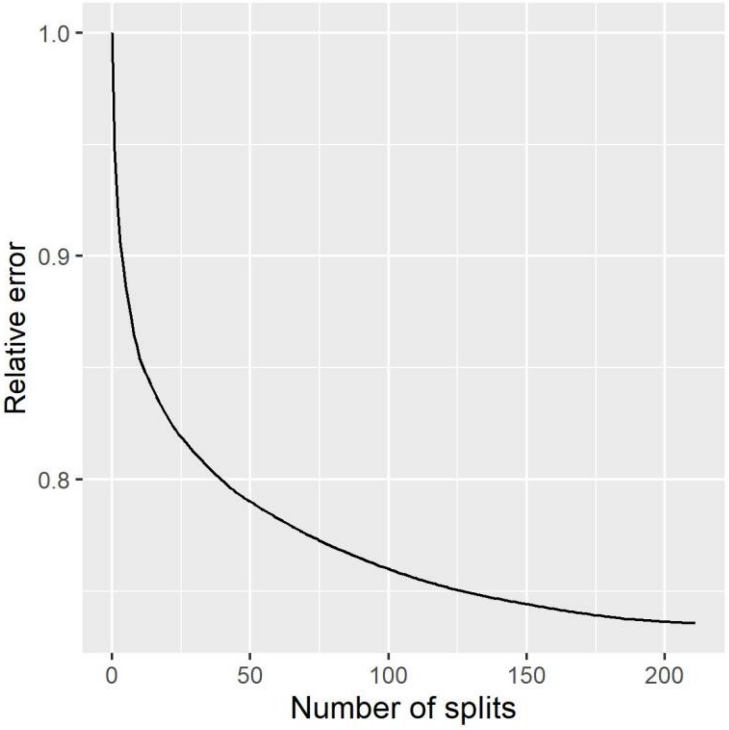
Source: Wage and Earnings Survey, own calculations

Fig 2. Validation error curve in 2009



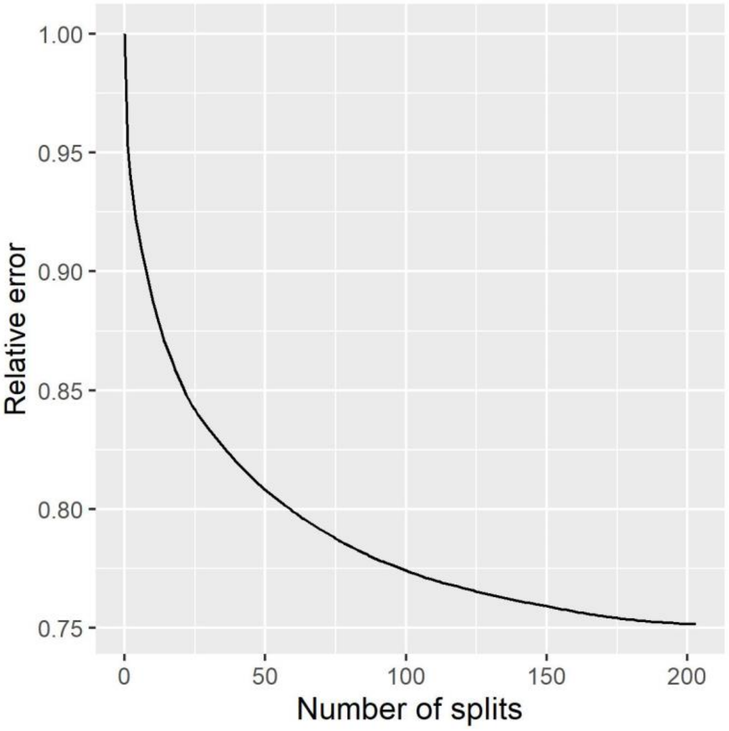
Source: Wage and Earnings Survey, own calculations

Fig 3. Validation error curve in 2010



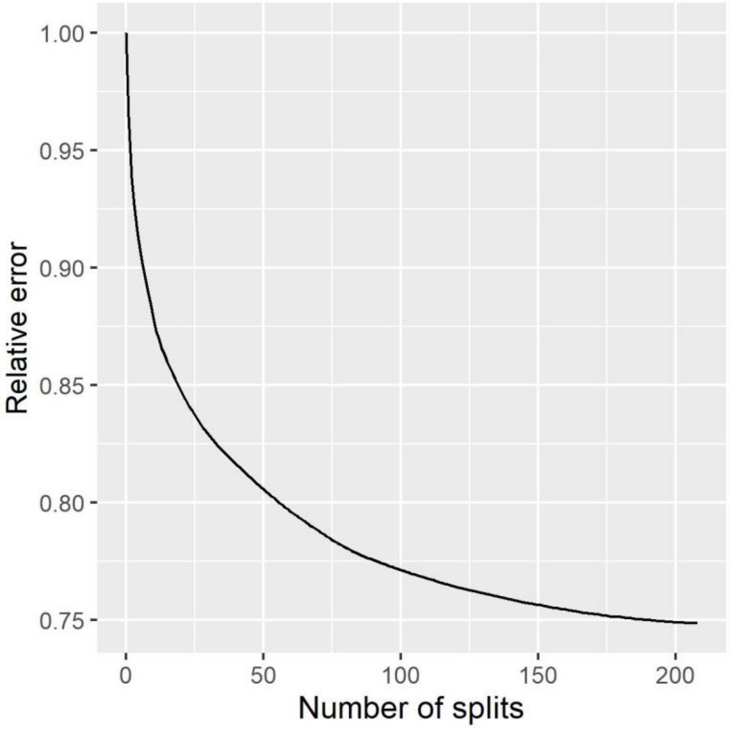
Source: Wage and Earnings Survey, own calculations

Fig 4. Validation error curve in 2011



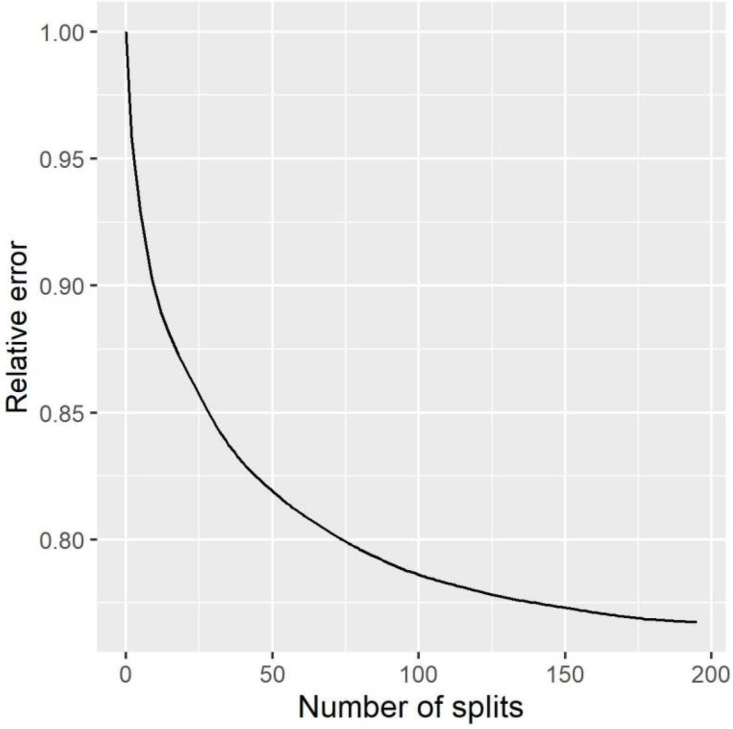
Source: Wage and Earnings Survey, own calculations

Fig 5. Validation error curve in 2012



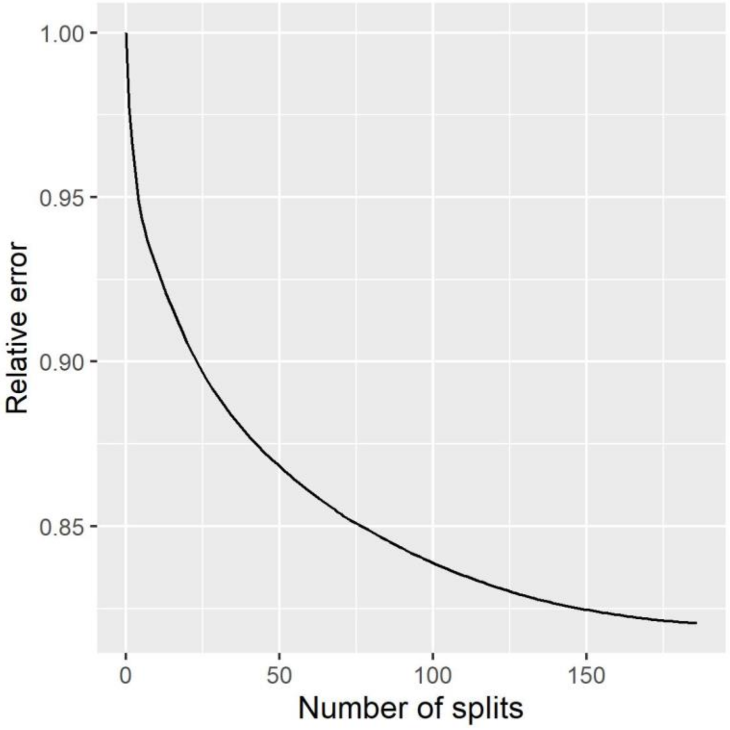
Source: Wage and Earnings Survey, own calculations

Fig 6. Validation error curve in 2013



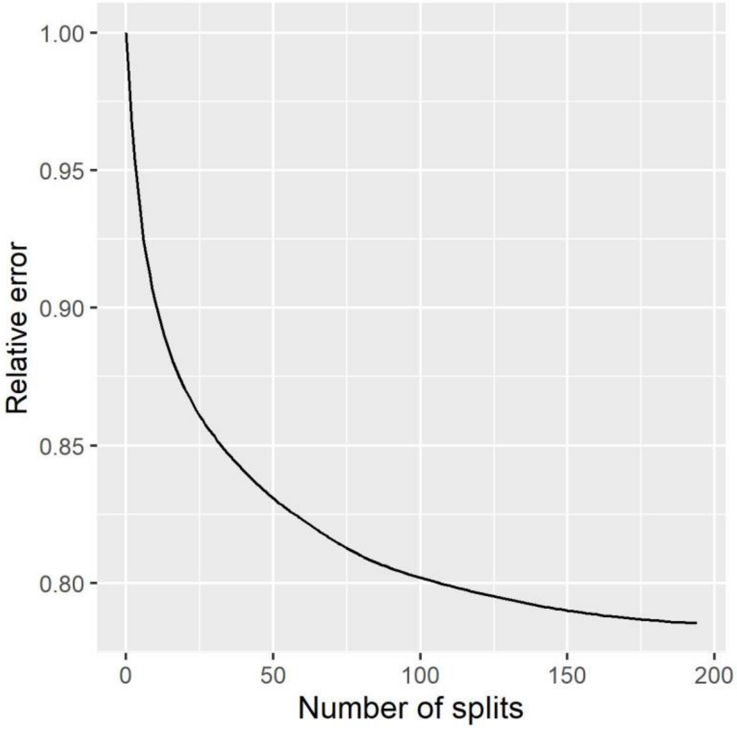
Source: Wage and Earnings Survey, own calculations

Fig 7. Validation error curve in 2014



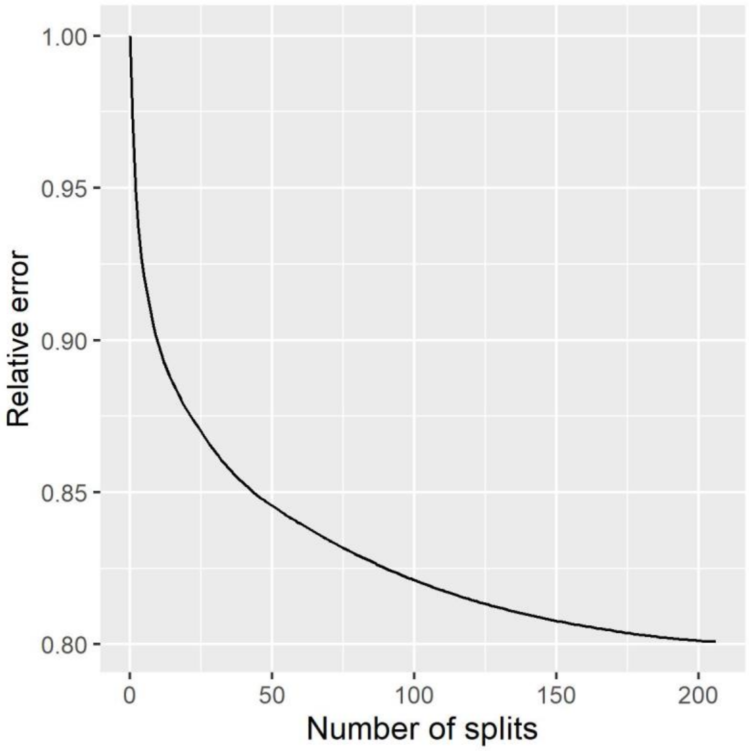
Source: Wage and Earnings Survey, own calculations

Fig 8. Validation error curve in 2015



Source: Wage and Earnings Survey, own calculations

Fig 9. Validation error curve in 2016



Source: Wage and Earnings Survey, own calculations

S4 Appendix H and L groups, 2008-2016

In Appendix S1 Tables 1-9 show the characteristics of the three groups with the smallest average WS effects, while Tables 10-18 those with the largest WS effects, as identified by the CART algorithms. Y denotes the average WS effect, Deviance is the within group mean squared error. Number is the number of observations in the group. The rest of the rows exhibit the implicitly defined restrictions on the corresponding covariates. (See the definition of covariates in Table 5.) An empty cell means that the covariate in question was not used in any splits.

Table 1. S-groups in 2008

	S1	***S2***	***S3***
Y	-0.025731	-0.0209026	-0.00957
Deviance	9.499419	26.53047	22.29703
Number	122	574	398
Sector	1, 3	4	4
Size	1	1	1, 2
Region	1	1	1, 3
Settlement		1, 3	
Foreign	3, 4	3, 4	3, 4
State			
Education	3, 4	2, 3, 4	
Tenure		1, 2, 3	4
Age	4, 5		
Agreement		0	

Source: Wage and Earnings Survey, own calculations

Table 2. S-groups in 2009

	S1	***S2***	***S3***
Y	-0.096688	-0.0176148	0.0251413
Deviance	25.34599	10.2645	14.54164
Number	377	192	369
Sector			1, 3
Size	1	1	1
Region	1	1	1
Settlement			
Foreign	4	4	4
State			
Education	4	4	3
Tenure	2, 3, 4	1	
Age			
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 3. S-groups in 2010

	S1	***S2***	***S3***
Y	-0.0305058	0.0360625	0.0374345
Deviance	16.31806	21.9751	39.28644
Number	352	540	1114
Sector	4	1, 3	4
Size	1, 2	1, 2	1, 2
Region			
Settlement		1, 3	
Foreign	4	4	4
State			
Education	3	3	3
Tenure	4		1, 2, 3
Age			
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 4. S-groups in 2011

	S1	***S2***	***S3***
Y	0.0276915	0.0376275	0.0591582
Deviance	63.93111	25.12144	18.66056
Number	1840	649	378
Sector	1, 3, 4	4	1, 3, 4
Size	1	2, 3	1
Region		1, 3	
Settlement			1, 3
Foreign	3, 4	3, 4	3, 4
State			
Education	3	3	4
Tenure			
Age			
Agreement			0

Source: Wage and Earnings Survey, own calculations

Table 5. S-groups in 2012

	S1	***S2***	***S3***
Y	-0.0141838	-0.00307359	0.0565023
Deviance	14.8207	10.06723	6.954345
Number	441	170	132
Sector	1, 3, 4	1, 3, 4	1, 3, 4
Size	1	1, 2	1,2
Region			
Settlement	1	1	2, 3
Foreign		3, 4	3, 4
State			
Education	3	4	4
Tenure			4
Age		1, 2	
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 6. S-groups in 2013

	S1	***S2***	***S3***
Y	0.00836586	0.0153548	0.0375331
Deviance	13.56098	8.080234	5.969932
Number	462	251	278
Sector	4	4	3
Size	1	2, 3, 4	1, 2
Region	1		
Settlement			
Foreign		1, 2, 3	
State			
Education	3	3	3
Tenure		1	
Age			
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 7. S-groups in 2014

	S1	***S2***	***S3***
Y	0.04594	0.052996	0.0580617
Deviance	38.41433	19.72705	47.25422
Number	887	764	2391
Sector	4	1, 3, 4	1, 3, 4
Size	2, 3, 4	2, 3	1
Region			
Settlement		1, 3	
Foreign	3, 4	3, 4	3, 4
State			
Education	1, 2, 3	1, 2, 3	1, 2, 3
Tenure			
Age	1, 2	3, 4, 5	
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 8. S-groups in 2015

	S1	***S2***	***S3***
Y	-0.0008278	0.0174181	0.0340221
Deviance	15.23211	33.56694	6.295836
Number	596	664	204
Sector			
Size	1, 2	1, 2	1, 2
Region	1	1, 3	
Settlement	3	1, 3	2
Foreign	2, 3, 4	3, 4	3, 4
State			
Education	2, 3	3, 4	3, 4
Tenure	1, 2	3, 4	3, 4
Age			5
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 9. S-groups in 2016

	S1	***S2***	***S3***
Y	0.0269456	0.0564116	0.0941928
Deviance	29.98993	7.439026	40.74489
Number	953	132	1770
Sector	3, 4		1, 2
Size	1, 2	1, 2	1, 2
Region	1		
Settlement			
Foreign	4	4	4
State			
Education	1, 2, 3	4	1, 2, 3
Tenure			1, 2
Age		5	
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 10. L-groups in 2008

	L1	***L2***	***L3***
Y	0.3585061	0.303635	0.2916854
Deviance	17.16966	9.070761	18.66253
Number	487	429	603
Sector	2	2	2
Size	1, 2	3, 4	4
Region	2, 3	2	
Settlement			
Foreign	1, 2, 3	4	1, 2, 3
State		2, 3, 4	
Education			1, 2, 3
Tenure			3, 4
Age			
Agreement			1

Source: Wage and Earnings Survey, own calculations

Table 11. L-groups in 2009

	L1	***L2***	***L3***
Y	0.3088706	0.3022631	0.2968929
Deviance	6.361225	5.473134	31.47635
Number	137	210	986
Sector	1	2	2
Size	4	2, 3	2, 3, 4
Region	2		2
Settlement			
Foreign		1	
State		3, 4	4
Education		1, 2	
Tenure		1, 2	3, 4
Age			
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 12. L-groups in 2010

	L1	***L2***	***L3***
Y	0.3798796	0.3316865	0.3080457
Deviance	17.3512	22.91723	8.817311
Number	550	375	330
Sector	2	1, 3, 4	2
Size	1, 2, 3		4
Region			
Settlement			
Foreign	1, 2, 3	1, 2, 3	1, 2, 3
State			
Education	1, 2	4	1, 2
Tenure	3, 4	1, 2, 3	3, 4
Age		3, 4, 5	
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 13. L-groups in 2011

	L1	***L2***	***L3***
Y	0.3942149	0.3343539	0.3336584
Deviance	7.353027	8.225938	9.33719
Number	218	222	316
Sector	2	1	2
Size	1, 2	4	3, 4
Region	2, 3	2, 3	2, 3
Settlement			
Foreign	1, 2	3, 4	1, 2
State			
Education	1, 2, 3		3
Tenure	3, 4		3, 4
Age			
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 14. L-groups in 2012

	L1	***L2***	***L3***
Y	0.3875253	0.3267591	0.3263497
Deviance	8.430231	5.698142	19.35548
Number	175	227	227
Sector	1, 3, 4	2	1
Size		3	3, 4
Region	2, 3		
Settlement			
Foreign	1, 2	3, 4	3, 4
State			
Education	4	2	4
Tenure		3, 4	
Age	3, 4, 5		
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 15. L-groups in 2013

	L1	***L2***	***L3***
Y	0.4005946	0.3030827	0.2980548
Deviance	8.113431	9.37924	8.518396
Number	189	318	193
Sector	2	1, 3	2
Size	3, 4	3, 4	
Region	2	2, 3	
Settlement			
Foreign			
State		3, 4	
Education	2	1, 2, 3	4
Tenure	4	3, 4	1, 2
Age			3, 4, 5
Agreement			

Source: Wage and Earnings Survey, own calculations

Table 16. L-groups in 2014

	L1	***L2***	***L3***
Y	0.309569	0.301662	0.2935012
Deviance	13.98765	9.510091	6.389392
Number	227	459	216
Sector	1, 4	2, 3	2, 3
Size	4	1, 2, 3	4
Region			
Settlement			
Foreign	1	1, 2	1, 2
State			4
Education	3, 4	1, 2	
Tenure	3	3, 4	4
Age	3, 4, 5		
Agreement			1

Source: Wage and Earnings Survey, own calculations

Table 17. L-groups in 2015

	L1	***L2***	***L3***
Y	0.3195566	0.3013315	0.2994339
Deviance	3.700527	4.970574	4.77308
Number	190	114	283
Sector	2	1, 3	2
Size	3, 4	2, 3, 4	1, 2, 3
Region		2, 3	
Settlement	2		
Foreign	3, 4		1, 2
State			
Education	1, 2, 3	4	1, 2
Tenure	4	1, 2	3, 4
Age			
Agreement	0		

Source: Wage and Earnings Survey, own calculations

Table 18. L-groups in 2016

	L1	***L2***	***L3***
Y	0.3204711	0.3110357	0.2875925
Deviance	8.721867	5.729796	4.171949
Number	220	173	254
Sector	1, 3	2, 3, 4	2, 3, 4
Size			1, 2, 3
Region			
Settlement		1, 3	2
Foreign	1, 2, 3	1, 2	1, 2
State			
Education	4	1, 2	1, 2
Tenure		3, 4	3, 4
Age	3, 4, 5	3, 4, 5	3, 4, 5
Agreement			

Source: Wage and Earnings Survey, own calculations