Takács, Olga; Vincze, János

Working Paper

The gender pay gap in Hungary: New results with a new methodology

IEHAS Discussion Papers, No. MT-DP - 2019/24

Provided in Cooperation with:
Institute of Economics, Centre for Economic and Regional Studies, Hungarian Academy of Sciences

Suggested Citation: Takács, Olga; Vincze, János (2019) : The gender pay gap in Hungary: New results with a new methodology, IEHAS Discussion Papers, No. MT-DP - 2019/24, Hungarian Academy of Sciences, Institute of Economics, Budapest

This Version is available at:
https://hdl.handle.net/10419/222069

INSTITUTE OF ECONOMICS
Centre for Economic and Regional Studies

KRTK KTI

Közgazdaság- és Regionális Tudományi Kutatóközpont
KÖZGAZDASÁG-TUDOMÁNYI INTÉZET

MŰHELYTANULMÁNYOK DISCUSSION PAPERS

MT-DP – 2019/24

The gender pay gap in Hungary:

new results with a new methodology

OLGA TAKÁCS – JÁNOS VINCZE

INSTITUTE OF ECONOMICS, CENTRE FOR ECONOMIC AND REGIONAL STUDIES,
BUDAPEST, 2019

Discussion papers
MT-DP – 2019/24

Institute of Economics, Centre for Economic and Regional Studies

The gender pay gap in Hungary: new results with a new methodology

Authors:

Olga Takács
Corvinus University of Budapest
and
Center for Economic and Regional Studies, Institute of Economics
email: takacs.olga@krtk.mta.hu

János Vincze
Corvinus University of Budapest
and
Center for Economic and Regional Studies, Institute of Economics
email: janos.vincze@uni-corvinus.hu

December 2019

The gender pay gap in Hungary: new results with a new methodology

Olga Takács – János Vincze

Abstract

We estimate the gender pay gap with the traditional OLS based Blinder-Oaxaca decomposition, and with an extension using Random Forest (RF) regressions on Hungarian data for the years 2008-2016. Random Forests perform better as predictors out-of-sample and yield consistently lower estimates for the unexplained pay gap than OLS. Then we analyse the unexplained gaps obtained from the RF regressions with a CART (Classification and Regression Tree) analysis. It seems that sectoral and educational factors are most consistently involved, but some other factors like firm size, age or tenure are also important. There are indications that medium educational levels and small firm size together, in certain industries, are most conducive to small (or even negative unexplained gaps), while high educational achievement in certain other industries (including manufacturing) are responsible for the highest gaps. In the first years of our sample period it was true in particular for middle aged and older women. This seems to be in accordance with the idea that educated women may have had problems with accumulating human capital.

JEL codes: J16 J31 C14

Keywords: gender pay gap, Hungary, Oaxaca-Blinder decomposition, Random Forest Regression, CART

A nemek közti béregyenlőtlenség Magyarországon: új eredmények egy új módszerrel

Takács Olga – Vincze János

Összefoglaló

A nemek közti bérkülönbségeket Magyarországon a hagyományos legkisebb négyzetek módszerével és a Véletlen Erdő regresszióval is megbecsüljük 2009–2016-os adatokon. Azt találjuk, hogy a Véletlen Erdő regressziók pontosabb becslést adnak, ezért a belőlük számított Blinder–Oaxaca-dekompozíciókat megbízhatóbbnak tekinthetjük. Ezek konzisztensen kisebb nem magyarázott bérkülönségeket adnak, amiket a potenciális nemi diszkrimináció mérőszámainak tekinthetünk Ezután a Klasszifikációs és Regressziós Fák segítségével elemezzük a nem magyarázott bérkülönbségeket. Úgy tűnik, hogy az ágazati és képzettségi tényezőknek van konzisztensen a legnagyobb súlyuk, habár más faktoroknak, mint a vállalatméretnek, az életkornak vagy a szolgálati időnek is fontos szerepe van. Vannak jelei annak, hogy a közepes képzettségi szint– bizonyos iparágakban – a kisvállalatokban jár leginkább együtt kis bérkülönbségekkel. A másik oldalon az egyetemi vagy főiskolai diplomások más iparágakban – így a feldolgozó iparban is – vannak kitéve a legnagyobb megmagyarázatlan bérkülönbségeknek. A mintaidőszak első részében ez különösen a középkorú és az idősebb nőket érintette. Ha ez valóban fennáll, akkor összhangban van azzal a nézettel, hogy a képzett nők számára az emberi tőke felhalmozása problémát jelentett.

JEL: J16, J31, C14

Tárgyszavak: nemek közti bérkülönbség, Magyarország, Oaxaca-Blinder-dekompozíció, Véletlen Erdő regresszió, Klasszifikációs és Regressziós Fák

The gender pay gap in Hungary: new results with a new methodology

Olga Takács

Corvinus University of Budapest and Center for Economic and Regional Studies, Institute of Economics

(takacs.olga@krtk.mta.hu)

János Vincze

Corvinus University of Budapest and Center for Economic and Regional Studies, Institute of Economics

(janos.vincze@uni-corvinus.hu)

Abstract

We estimate the gender pay gap with the traditional OLS based Blinder-Oaxaca decomposition, and with an extension using Random Forest (RF) regressions on Hungarian data for the years 2008-2016. Random Forests perform better as predictors out-of-sample and yield consistently lower estimates for the unexplained pay gap than OLS. Then we analyse the unexplained gaps obtained from the RF regressions with a CART (Classification and Regression Tree) analysis. It seems that sectoral and educational factors are most consistently involved, but some other factors like firm size, age or tenure are also important. There are indications that medium educational levels and small firm size together, in certain industries, are most conducive to small (or even negative unexplained gaps), while high educational achievement in certain other industries (including manufacturing) are responsible for the highest gaps. In the first years of our sample period it was true in particular for middle aged and older women. This seems to be in accordance with the idea that educated women may have had problems with accumulating human capital.

JEL codes: J16 J31 C14

Keywords: gender pay gap, Hungary, Oaxaca-Blinder decomposition, Random Forest Regression, CART

1 Introduction

Wage discrimination is a paramount social problem, and it is also challenging from an intellectual point of view. Obviously, recording a non-zero difference between the earnings of two distinct groups cannot be taken as evidence for unequal treatment, and though a purely statistical approach could not give a valid verdict on discrimination, it could greatly help to refine our understanding of available facts. In this paper, focusing on the gender wage gap, we wish to contribute to this refinement by applying a novel methodology that may be regarded as improved observation.

The econometric literature on wage discrimination based on gender is huge. According to an OECD study the gender wage gap has recently declined worldwide, but it still exists, as women's income was still 39 percent lower than men's in 2015. (OECD (2018)) According to Weichselbaumer–Wintner-Ebmer (2006), the number of papers dealing with this topic has been growing exponentially since the 1990s. The existence of the gender wage gap is a hot topic not only in labour economics, but, also, a policy priority in the EU and in Hungary, as well. For example, the European Union has a plan "Strategic engagement for gender equality 2016-2019". In parallel with this strategic engagement member states, including Hungary, have their own action plans[1] (European Commission (2015)). Indeed, the principle of "equal pay for equal work" can be found in the Hungarian labour code.

For the empirical assessment of the gender wage gap the most frequently applied methodology is the Blinder-Oaxaca decomposition (see (Darity-Mason (1998)). This decomposition is a general empirical methodology (see Jann (2008) for a general review) for breaking down the difference between the means of two sub-populations into a part, explainable by observed features, and the rest (the "unexplained" part). In the original two-way decomposition framework one specifies and estimates three linear regression models: one for each subpopulation and a reference model, possibly for the whole population. To identify the unexplained part with discrimination the reference model should be structural, derivable from some theory of wage determination without discrimination. Frequently, it is the human capital theory, and the reference model is a Mincer-equation. In theory, the pure discrimination interpretation of the unexplained part amounts to differences in regression coefficients. In

[1] In Hungary it is the Government Resolution No. 1004/2010 (I. 21.) National Strategy for the Promotion of Gender Equality - Guidelines and Objectives 2010-2021.

practice, however, even if one had a well-established empirical model for wage determination the estimates would probably be inconsistent, due to selection bias, unobserved heterogeneity and errors-in-variables problems. (Kunze (2008) pp. 63-76.) Even if one thinks that gender is randomly assigned it is not obvious whether the differences between regression coefficients stem from unobserved heterogeneity or discrimination. As non-experimental data alone cannot answer this issue, Neuman and Oaxaca (2004) conclude that value judgement is necessary with respect to what constitutes discrimination. Analysis of the literature (Weichselbaumer–Wintner-Ebmer. (2006, pp. 416-436.) shows that many researchers have preferred the term „unexplained residual" rather than interpreting results as bearing on the issue of discrimination. From this point of view, the underlying models do not express structural relationships, they can be considered as predictions based on some information set, and the decomposition is best regarded as a descriptive device.

The original Blinder-Oaxaca framework involves linear regression analysis. In this paper we emulate the spirit of the Blinder-Oaxaca decomposition, and compute explained and unexplained parts on Hungarian data. However, rather than using OLS estimates we compute the decomposition via Random Forest (RF) regressions. The methodology and its justification is detailed in the Takács-Vincze (2019). Besides estimating the pay gap, we analyse the unexplained parts with CART models, offering hints of which parts of the population are most and least exposed to "unexplained" gender pay differences.

In Section 2 we overview the related literature with particular focus on previous Hungarian results. Section 3 contains the methodology, data and results, while Section 4 concludes.

2 Related literature

2.1 The gender pay-gap in general

Conducting a large scale meta-analysis of the gender wage gap literature Weichselbaumer–Winter-Ebmer (2005) uncovered a few general features. 1. The wage gap had decreased over time. 2. The lack of good human capital variables exaggerates the gap. 3. Narrow group investigations (restricting attention to some subpopulation, for instance some occupational group), reveal smaller unexplained gaps. 4. The concrete econometric methodology does not matter too much.

Applying the Blinder-Oaxaca methodology researchers invariably found that the raw gender wage gap is positive (i.e. men's average wages are higher), and the unexplained gap is positive, too. (See Darity-Mason (1998)) However, there are differences with respect to the sign of the explained part. For several countries it is negative, it seems as if in a non-discriminatory regime women's average wages must be higher than that of men's. These countries usually are South and Eastern European ones, with a relatively low female employment rate. (See Olivetti–Petrongolo (2008) and Leythienne–Ronkowski (2018)) Also Leythienne–Ronkowski (2018) found strong occupation selection effects: more women work in low-paying occupations. These findings induced the question whether the wage regression should be complemented with a binary participation regression, and the parameters of the wage regression adjusted accordingly.

In connection with the gender wage gap OECD (2018) emphasizes the consequences of childbirth on female career paths. Women tend to stay at home with children resulting in less work experience, and missing career opportunities that lead to lower income.[2] Furthermore, a job can have such non-financial features as unstressed workplace, less competitiveness or more flexible hours which may accord better with the needs of parenting. However, these features can produce self-sorting in the labor market of a kind that women end up in lower paid jobs than men. Thus, the distribution of women and men are not equal across occupations. Blau-Kahn (2000) and Hegewisch-Hartmann (2014) found this kind of segregation reduced by the early 2000s, but it still exists in the USA. Accounting for the sorting of women into lower paying jobs attempts to make the unexplained part strictly a measure of wage-setting discrimination. (Simón (2012)), Drolet-Mumford (2012), Bayard et al. (2003), Card-Cardoso-Kline (2015)) Indeed, in some cases, it turned out that estimates would change significantly after a Heckman-correction (see for example Gannon et al (2007)).

As mentioned in the Introduction there can be observed a tendency to deemphasize the structural nature of the wage regression. This led to the inclusion of variables that are not of the human capital type, such as industry dummies, firm specific variables, or the femininity of the occupation (see Bayard et al. (2003)). In general, the outcome was a narrowing of the unexplained gap. As researchers introduced more and more variables regarding firm and job characteristics they developed new theories about the unexplained part. Some of them contended that bargaining institutions matter, and that the gender gap is partly due to poorer

[2] For a detailed analysis about the connection of gender wage gap and motherhood in Hungary, see Cukrowska-Lovász (2014).

bargaining strength of women (Gannon et al. (2007)). Others emphasised the effect of childbearing and caring on women's decisions (e.g. missing promotion opportunities, wish for flexible working time) which can cause self-selection into lower paying jobs and/or less experience (Cukrowska–Lovász (2014), OECD (2018)).

In another line of research unobserved heterogeneity was emphasized, and demographic variables, like family size, or effects due to subject of degree were taken into account, making the unexplained gap smaller (see Machin-Puhani (2003)). Altonji and Blank (1999) in the USA, and Cheng (2005) on Canadian data, argued that controlling for occupation decreased the unexplained part of the regression. In contrast Kee (2005) and Barón and Cobb-Clark (2008) found that the Australian unexplained gender wage gap increased when they used occupation and industry variables. Chatterji et al. (2007) had similar results in Britain. So, results from the literature are contradictory.

The pay gap may differ between the public and private sectors. Lovász (2013) argued that positive nonfinancial features of jobs are present more distinctively in the public, than in the private sector, and this can be the cause of why working in the public sector is attractive for women. Unfortunately, these non-financial features are not easily measured. Other things being equal this would increase the pay gap. However, public sector wages are determined by political decisions in public sector while in private sector wage setting is affected by market conditions (Gregory–Borland (1999)), and anti-discrimination legislation may be more aggressively enforced in the public sector (Barón–Cobb-Clark (2008), pp. 2). The net effect is that the wage gap is higher in the private sector, as it is documented by Greene-Hoffnar (1996) in the USA, by Barón and Cobb-Clark (2008) on Australian data, by Chatterji et al. (2007) for Britain, by Cheng (2005) for Canada, and by Lovász (2013) for Hungary.

2.2 The gender pay-gap in Hungary

The gender wage gap has a decreasing trend in Hungary, and it is one of the smallest within the OECD. At the beginning of the 1990's market reforms led to a non-centralized wage-setting system in the private sector (Jolliffe–Campos (2005), Newell–Reilly (2001)). These reforms and other changes reduced the gender wage gap in Hungary, and, according to Brainerd (2000), the decrease in the total gap was due mainly to the narrowing of the "explained" gap, women's wages increased more than males' with similar characteristics. In other words, during and after

transition women's characteristics were revalued. Leythienne and Ronkowski (2018) found that the explained part is slightly negative.

There have occurred changes in the "unexplained" part, too. Blinder-Oaxaca type evidence for Hungary has been provided by Jolliffe–Campos (2005). This paper intended to measure the change in discrimination, and concluded that it was reduced over time, especially in big firms. Lovász (2008) argued that during transition domestic and international competition increased, which made discrimination costly, and it was another cause of the narrowing of the gap. Csillag (2007) underlined the role of occupational segregation. As relative wages stagnated occupational segregation diminished in the second part of the 90's. It seemed that a simple linear regression cannot account for every piece of the developments. Gábor (2008) discovered interesting non-linearities in the earning function. He examined the effects of human capital variables (experience, educational level and gender) on earning age-profiles between 1992 and 2003, and found that men's earning profiles start higher than women's at every educational level, and, up to a certain point, they are steeper, due, probably, to accumulating more experience earlier in life. This suggests that age must be an important non-linear factor in explaining wages, and also that if we cannot observe work experience it may give a boost to the unexplained gap.

3 Methodology, data and results

3.1 Methodology

The original Blinder-Oaxaca framework involves linear regression analysis. This has the additional advantage of enabling investigators to decompose the unexplained and explained parts variable-wise, too. (See (Gardeazabal-Ugidos (2004)) for some difficulties, and a proposed solution.) However, some types of problems require non-linear modelling (e.g. binary outcomes and Tobit models), and the Blinder-Oaxaca methodology has been duly generalized in that direction. (See Fairlie (2005), Bauer-Sinning (2010).) Herein we apply an extension of the Blinder-Oaxaca decomposition outside the traditional regression framework. With the help of a specific statistical learning algorithm, Random Forest regression, we follow the main lines of the Blinder-Oaxaca decomposition, and compute Blinder-Oaxaca style decompositions for the gender wage gap on Hungarian data. The traditional methodology produces variable-wise decompositions, which is not available with ours. Rather than attributing the gaps to individual variables we carry out a CART (Classification and Regression Tree) analysis to segregate the

samples into “audiences”, i.e. to discover those subgroups that are more or less responsible for the unexplained gap.

We employ the methodology described in Takács-Vincze (2019). This consists of the following steps. 1. We estimate three RF regressions on a training sample: a reference model, which in our case is an RF regression without the gender variable, a male and a female model on the respective subsamples. 2. Then we predict female and male wages with the reference model, and taking the difference between the predicted means we obtain the explained gender wage gap. 3. Then we calculate individual prediction differences between the reference prediction and the own prediction for both females and males. The difference between these differences gives the unexplained pay gap individually. The average of these individual pay gaps gives us the unexplained gap. 4. We take the individual pay gaps for females and males separately, and run a CART on each. These CARTs offer us classifications concerning which sections of the society exhibit the highest and lowest degree of the unexplained pay gap.

We checked the reliability of our methodology in two ways. As with RF regressions the explained and unexplained gaps do not sum up to the raw gender pay gap we calculated the difference, which we call the bias. Second, we needed to justify the application of the tree-based methodology by comparing the out-of-sample prediction power of RF with that of OLS.

3.2 Data

The dataset used in this study is the Hungarian Wage Survey Data, hosted by the National Employment Office. It is a matched employer-employee database that provides annual information (recorded for May each year) on workers’ age, gender, occupation, earnings (disaggregated into regular pay and irregular bonuses), type of contract, and whether the worker was hired recently. It also contains information about the employer (sector, region, settlement type, size of employment, etc.). The sampling procedure of corporate employees is based on firm size. Each annual sample includes all firms with more than 50 employees and a randomly sampled part of firms with 5-50 employees.

We have used the logarithm of the gross monthly wage for our analysis. The gross monthly wage includes regular wage and bonuses as well. As this variable is inappropriate to compare full-time and part-time employees, and also not very useful for including public employees, we restricted our sample to employees working full-time in the private business sector. Our

calculations were carried out for a training and a test sample for each year between 2008 and 2016. The training samples contained 60 000 annual observations, and the rest made up the test sample. Table 1 reports basic information about the samples.

Table 1 Number of observations and the raw pay gap in the dataset

Year	Number of observations	Raw gap (wages in logarithm)
2008	131 922	0.0897
2009	119 944	0.0670
2010	120 432	0.0937
2011	119 980	0.0980
2012	121 690	0.1140
2013	125 103	0.1095
2014	136 343	0.1123
2015	156 338	0.1204
2016	134 704	0.1164

Source: Hungarian Wage Survey (citation to KRTK Data Base)

Unfortunately, we do not observe employee characteristics like true work experience, marital status, or subject of degree, variables that would lend a strong human capital interpretation of the earning function. However, as our main goal is descriptive rather than explanatory we include many variables that could help predict wages, even if they may be endogenous. Table 2 lists the predictors that we used in both the OLS and RF regressions. Notice that in the OLS regressions age-squared was also included.

Table 2

List of variables (in parentheses the corresponding names in the charts)

Name	Unit
Age (kor)	Years
Tenure (szolgho)	Months (at current employer)
Education (iskveg9_ordered)	9 levels (ordinal) (levels 8 and 9 are College and University)
New entrant dummy (ujbel)	0: no, 1: yes
Share of foreign property (kra_ordered)	4 levels (ordinal), 1: 100 % foreign property
Share of state property (ara_ordered)	4 levels (ordinal) 1: 100 % state-owned
Firm size (letszam_bv1)	Number of employees
Settlement (ttip)	1:Capitalcity(Budapest),2:Town, 3: Other
Region (kshreg)	7 categories
Sector (ag1)	NACE Rev. 2 - 2 digits
Collective agreement on firm level (kol)	0: no, 1: yes
Collective agreement on sector level (kag)	0: no, 1: yes
Collective agreement with more employers but not on sector level (ksz)	0: no, 1: yes
Share of white collar employees in enterprise (szesu_v1)	percentage

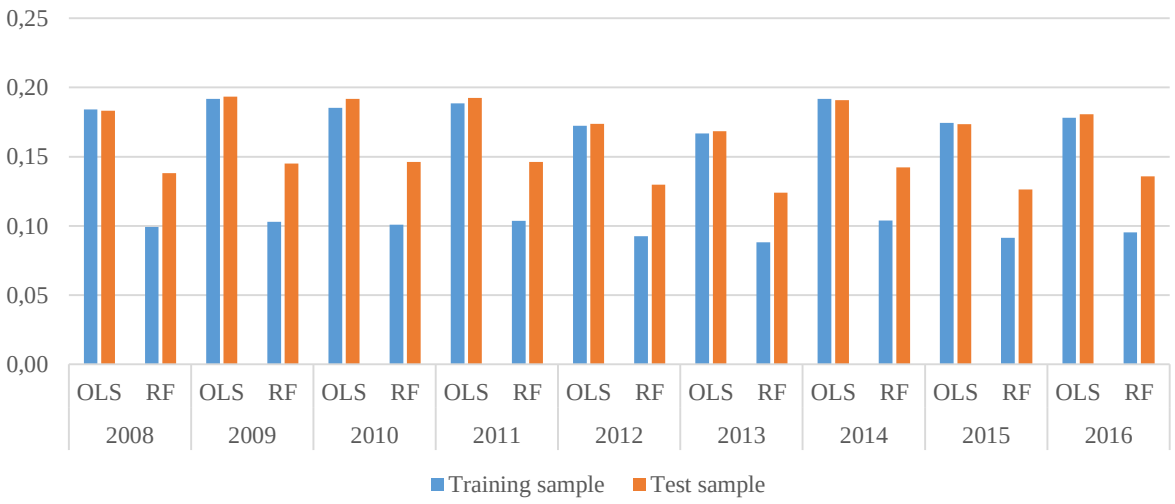
Source: Hungarian Wage Survey

The Random Forest algorithm requires the setting of several parameters. In particular, our forests contained 500 trees each. To control for the growth of individual trees we set the minimum node-size parameter at 5, and we did not limit the maximum number of nodes. At every node the number of randomly selected variables was 5 (out of 14 explanatory variables).

3.3 Results

In Figure 1 columns show mean squared errors (MSE) for log wages based on training and test sample predictions. For all years the Random Forest model has smaller MSE than the OLS, both in the training and the test samples, though the difference is larger in the training samples.[3] Thus the Random Forest Regression seems to be a superior data description tool in our case.

Figure 1 Comparison of predictive performance of OLS and Random Forest

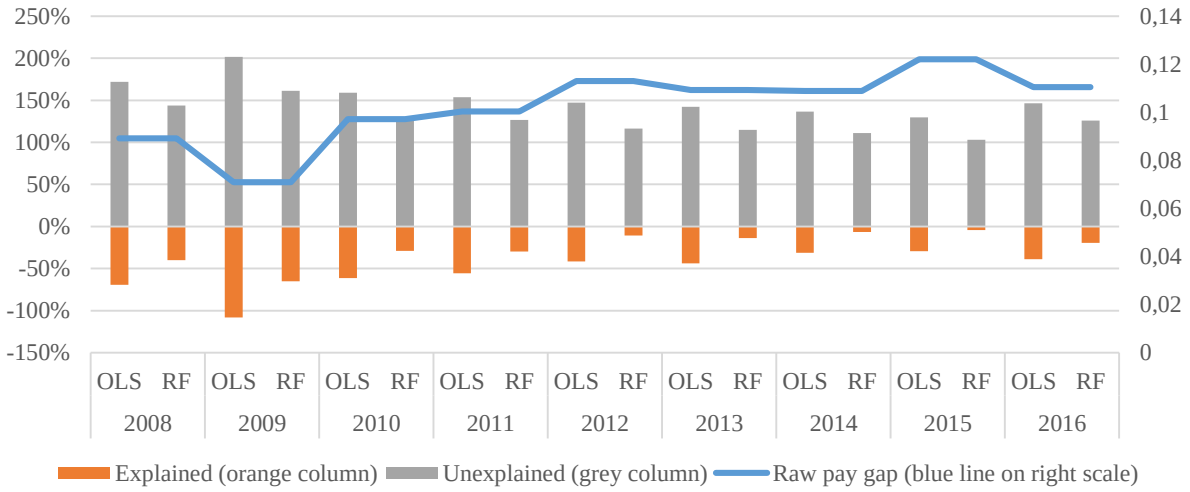


Source: Hungarian Wage Survey, own calculations

The raw gaps in log points, and their decompositions into explained and unexplained parts as percentages are presented in Figure 2 for the test samples. (Maximum biases are 8% with OLS, and 6% with Random Forest Regression.) The unexplained gaps are always smaller with the Random Forest estimates, while the explained gaps are negative in both cases. Thus the Random Forest Regressions suggest smaller effects to the difference between gender-dependent wage-setting mechanisms. The neutral (i.e. reference) estimates predict higher average wages for women than for men, in accordance with the former literature.

[3] Random Forest Regressions were calculated by the randomForest package in R (https://www.stat.berkeley.edu/~breiman/RandomForests/).

Figure 2 Elements of Oaxaca-Blinder decomposition



Source: Hungarian Wage Survey, own calculations

Figure 3.1 in the Appendix shows the optimal "small tree" obtained by setting the complexity parameter at 0.01. We have six leaves of which the leftmost exhibits the smallest unexplained gap. Women working in most sectors (except 3,4,6 and 10) and having no higher education degree belong here (46 % of the sample). The second smallest average gap is exhibited in a group where the properties of the lowest educational levels (1,2) and of the remaining sectors (3,4,6 and 10) are combined (7%). At the other extreme there are individuals from the latter sectors with education above the lowest and aged above 34.

From Figure 3.2 it seems that in 2009 the first sectoral cut causes a somewhat different sectoral division, now Sectors (2,3,4) are separated from the rest. The smallest unexplained gap appears in a group representing 6% of the sample with women working in many sectors, having higher educational achievement and being younger than 31. The leftmost group has now the second lowest unexplained gap, and is very similar to the leftmost group in 2008, also it is the largest (55%). At the other extreme the group showing the largest gap is similar to the corresponding group in 2008, but here there is no age restriction, and the sectoral cut is somewhat different.

The Figure for 2010 is broadly consistent with the former ones. The smallest gap is exhibited by a large group (51 %) with no higher education and belonging to a wide range of sectors, while the second lowest (there is a slight difference) by a smaller group (6 %) with higher

degrees in the same sectors and with an age below 33. The largest gap shows up on the rightmost again, where the sectors include 3, 4, 5 and 10, education is above the most basic, and, for a novelty, firms have more than 79 employees.

For 2011 some changes can be observed. The largest (leftmost) group shows the smallest gap again. To this group belong most sectors (again with a small variation in the exact list), and women with no higher education performance. Now most other groups have medium gaps, while again the rightmost group (23 %) features the highest gap. Four cuts were needed to make this group: 1. sectoral (practically the usual suspects, including 3 and 4), 2. education (above the lowest), 3 firm size (above 31), and age (above 28).

Concerning the smallest gap the year 2012 almost copies the year 2011 (see Figure 3.5). With respect to the largest gap the rightmost extreme reproduces the former cuts with respect to sector and education, but now a new variable enters: the share of white collar employees in the firm. The cut selects for the rightmost group the firms above 4 of that indicator, which means above median but below average.

For 2013 consistency again prevails for the group with the smallest gap. For the largest gap the sectoral cut remains the same, the educational divide reappears, the white collar share is there somewhat loosened (now the group contain more than the median), and a new variable shows up: tenure (tenure above 53 practically means above median).

Further consistency can be observed on Figure 3.7. On the leftmost side it is almost perfect, on the rightmost side there is now three cuts: 1. The usual sectoral, 2. The usual educational, 3. Now it is the white collar cut (again above median and below mean).

The 2015 tree is very simple. The leftmost leaf is the usual. The rightmost now contains fewer cuts and more observations. The two cuts are: 1. the ordinary sectoral, and 2. a broad white collar cut (above median).

The 2016 tree is an almost exact copy of the 2010 tree. The leftmost group with the smallest gap has two cuts: 1. the usual sectoral and 2. the usual educational. The rightmost group with the largest gap has three cuts: 1. the usual sectoral, 2 the most frequent educational and 3. a firm size cut (above median but much below the mean).

It seems that the lowest levels of the gap appear for women who work in most sectors and do not have higher education qualifications. This group makes up about 50 % of the sample, A significant but much smaller group exhibits high gaps. This group features a select class of sectors, sector 3 and 4 are always among them. The other lines of division are not so clear, but it seems as if education above the basic level were a selection criterion, as well as working in not too small firms with an above median share of white collar employees, and having above median tenure. The role of age is not clear, but being young seems to be not associated with having a large gap, and its effects are apparently faded through time, as only the first years in the sample exhibit age as a cut variable.

Next we ran CART regressions where the complexity parameter was set at 0.0005, but we restrained tree growth by requiring that smallest leaf contains at least 50 observations. The resulting trees were very large, containing 88 leaves, for instance, for 2008.

We selected the three groups with the smallest gap, where the gap is actually negative. In other words, women belonging to these groups are un-explainedly well-paid. Sectorally and educationally these groups seem to be refinements of the lower tail groups in the small trees. The most interesting new group is the one with university education, small firms and relatively aged workers. Apparently, working in small firms gives a better chance to have small gaps.

Looking at the upper tail we find more or less the same sectors as in the small tree, and educational levels turn out to be high. Ages are close to middle, and small firms are absent. In the highest gap group fully foreign owned enterprises are also present. For the year 2009 the lower tail shows roughly the same picture as the 2008 tree with respect to sectors and education. However, other variables wildly differ among sectors. In the upper tail enterprise size and education behave similarly to the small trees, but sectorally the dispersion is high. The large trees for other years (not reported) show rather noisy pictures. The lower tails are generally consistent with the former sectoral divisions, and also with the educational one, medium levels of education dominate. In the upper tails the usual suspects, sectors 3 and 4 reappear, while the educational levels are rather high. Firm sizes also tend to be medium to high, but it is not unequivocal. In fact, the large trees did not provide interesting supplementary information, they show rather high volatility with respect to classification. Therefore, we did not pursue this line of investigation further.

4 Conclusion

Why is Random Forest Regression an attractive alternative to the traditional OLS based methodology? We believe that a model with a more accurate out-of-sample performance is a better approximation of the conditional expectation function. If we have a good statistical learning model that outperforms OLS in this respect, then we get a more reliable diagnosis, though without being able to prescribe a cure. Our findings suggest that observable characteristics explain more of the gender wage gap than traditional OLS estimates imply.

We tried to make sense of the individually estimated pay-gaps by a CART analysis. It seems that sectoral and educational factors are most consistently involved, but some other factors like firm size, age or tenure are also important. The CART methodology is not appropriate for separating the influence of individual variables, but it can be used for detecting interactions and non-linearities. There are indications that medium educational levels and small firm size together, in certain industries, are most conducive to small (or even negative unexplained gaps), while high educational achievement in certain other industries (including manufacturing) are responsible for the highest gaps. In the first years of our sample period it was true in particular for middle aged and older women. This seems to be in accordance with the idea that educated women may have had problems with accumulating human capital (see Csillag (2007)).

5 References

Altonji, J. G.–Blank, R. M. (1999): Race and Gender in the Labor Market. In: Ashenfelter, O.–Card, D. (edit.): Handbook of Labor Economics. Elsevier, Amsterdam, Vol. 3, pp. 3144–3259

Barón, J. D., & Cobb-Clark, D. A. (2010). Occupational segregation and the gender wage gap in private-and public-sector employment: a distributional analysis. *Economic Record*, *86*(273), 227-246.

Bayard, K., Hellerstein, J., Neumark, D., & Troske, K. (2003). New evidence on sex segregation and sex differences in wages from matched employee-employer data. *Journal of labor Economics*, *21*(4), 887-922.

Blau, Francine D., and Lawrence M. Kahn. "Understanding international differences in the gender pay gap." *Journal of Labor economics* 21.1 (2003): 106-144.

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

Card, David, Ana Rute Cardoso, and Patrick Kline. "Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women." *The Quarterly Journal of Economics* 131.2 (2015): 633-686.

Cukrowska, Ewa–Lovász, Anna (2014): Are children driving the gender wage gap? Comparative evidence from Poland and Hungary. Budapest Working Papers on the Labour Market BWP, issue 4

Csillag, Márton (2007): „Female work" and the gender wage gap form late socialism to today. In Galasi, Péter and Gábor Kézdi (editors): The hungarian labour market review and analysis, Budapest, Institute of Economics

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

Drolet, Marie, and Karen Mumford. "The gender pay gap for private-sector employees in Canada and Britain." *British Journal of Industrial Relations* 50.3 (2012): 529-553.

Elder, Todd E., John H. Goddeeris, and Steven J. Haider. "Unexplained gaps and Oaxaca–Blinder decompositions." *Labour Economics* 17.1 (2010): 284-290.

European Commission (2015): Strategic engagement for gender equality 2016-2019. Luxembourg, Publications Office of the European Union

Gannon, B., Plasman, R., Ryex, F., & Tojerow, I. (2007). Inter-industry wage differentials and the gender wage gap: evidence from European countries. *Economic and Social Review*, *38*(1), 135.

Gardeazabal, Javier, and Arantza Ugidos. "More on identification in detailed wage decompositions." *Review of Economics and Statistics* 86.4 (2004): 1034-1036.)

Jolliffe, Dean, and Nauro F. Campos. "Does market liberalisation reduce gender discrimination? Econometric evidence from Hungary, 1986–1998."*Labour Economics*12.1 (2005): 1-22.

Kunze, Astrid. "Gender wage gap studies: consistency and decomposition." *Empirical Economics* 35.1 (2008): 63-76.)

Leythienne, D. E. N. I. S., and P. I. O. T. R. Ronkowski. "A decomposition of the unadjusted gender pay gap using Structure of Earnings Survey data." *Luxembourg, Publications Office of the European Union. doi* 10 (2018): 796328.

Lovász, Anna (2013): Jobbak a nők esélyei a közszférában? A nők és férfiak bérei közötti különbség és a foglalkozási szegregáció vizsgálata a köz- és magánszférában. Közgazdasági szemle, vol. 60, pp. 814-836

Machin, Stephen, and Patrick A. Puhani. "Subject of degree and the gender wage differential: evidence from the UK and Germany." *Economics Letters* 79.3 (2003): 393-400.

Newell, Andrew, and Barry Reilly. "The gender pay gap in the transition from communism: some empirical evidence." *Economic Systems* 25.4 (2001): 287-304.

Neuman, Shoshana, and Ronald L. Oaxaca. "Wage decompositions with selectivity-corrected wage equations: A methodological note." *The Journal of Economic Inequality* 2.1 (2004): 3-10.)

Neumark, D., 1988. Employers' discriminatory behavior and the estimation of wage discrimination. The Journal of Human Resources, 23(3), 279-295. http://dx.doi.org/10.2307/145830

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

Simón, Hipólito. "The gender gap in earnings: an international comparison with European matched employer–employee data." *Applied Economics* 44.15 (2012): 1985-1999.

Takács, Olga, János Vincze (2019) Blinder-Oaxaca decomposition with recursive tree-based methods: a technical note, Budapest, CERS Working Papers, 23/2019.

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

Wager, S., & Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*.

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

Weichselbaumer, Doris, and Rudolf Winter-Ebmer. "A meta-analysis of the international gender wage gap." *Journal of Economic Surveys* 19.3 (2005): 479-511.

6 Appendix

Figure 3.1 CART for 2008

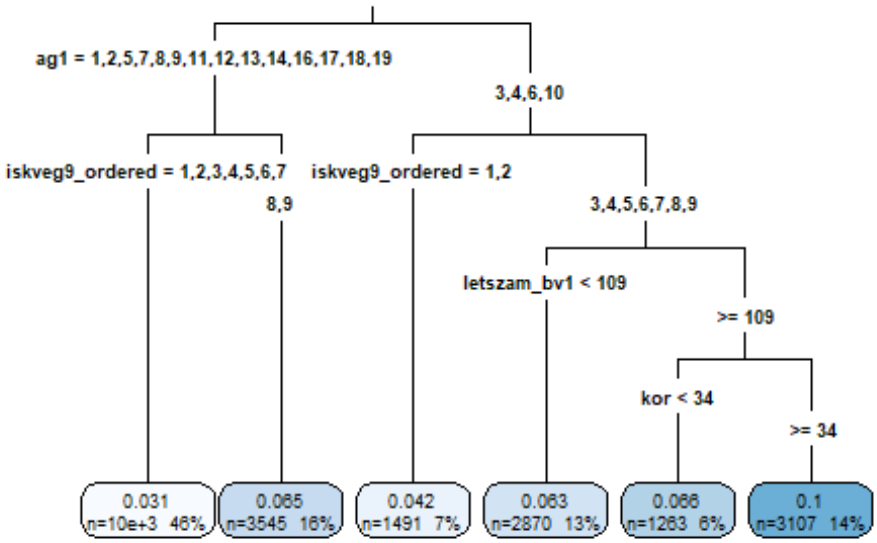


Figure 3.2 CART for 2009

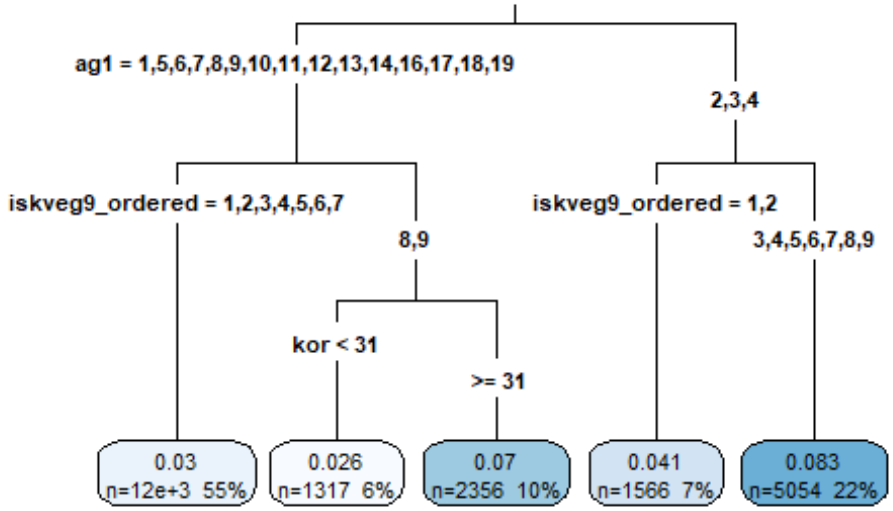


Figure 3.3 CART for 2010

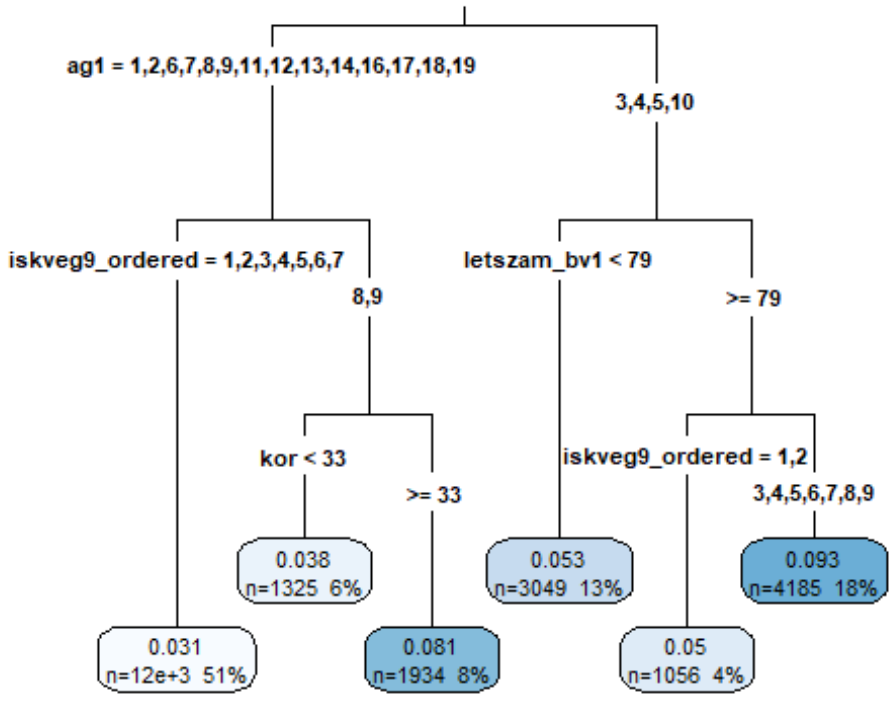


Figure 3.4 CART for 2011

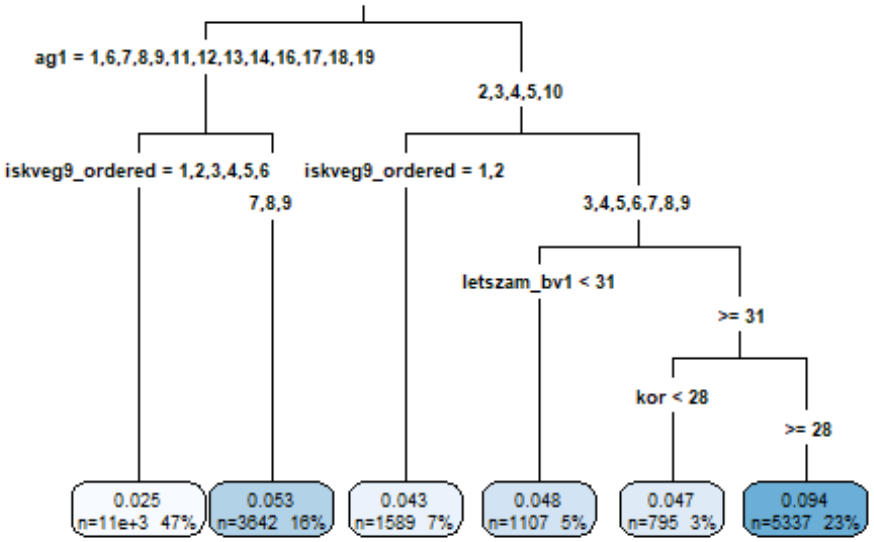


Figure 3. 5 CART for 2012

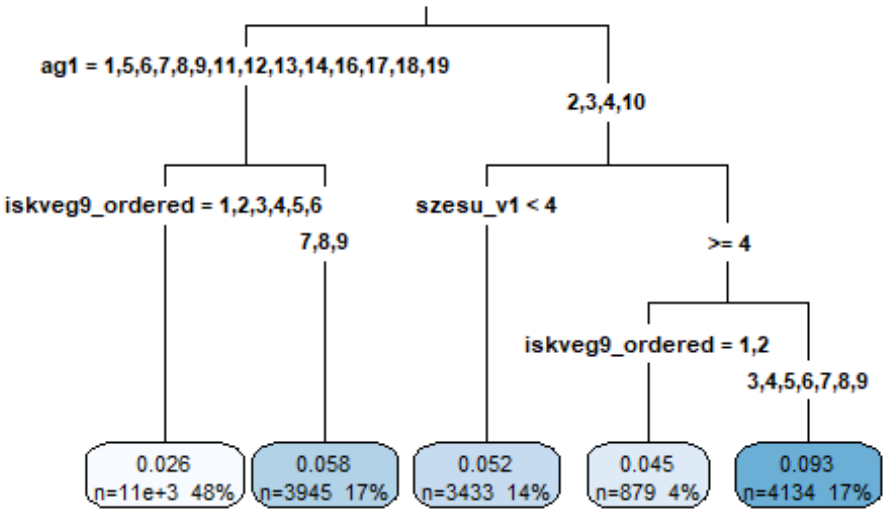


Figure 3.6 CART for 2013

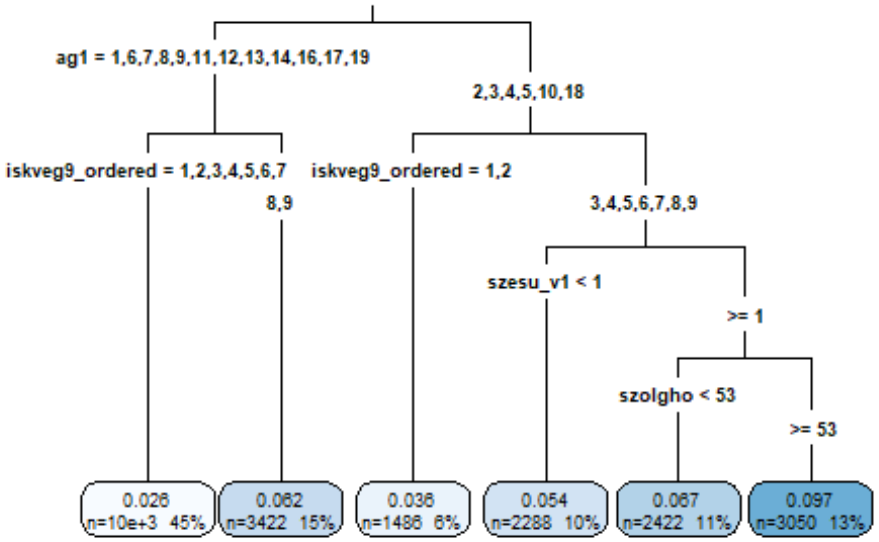


Figure 3.7 CART for 2014

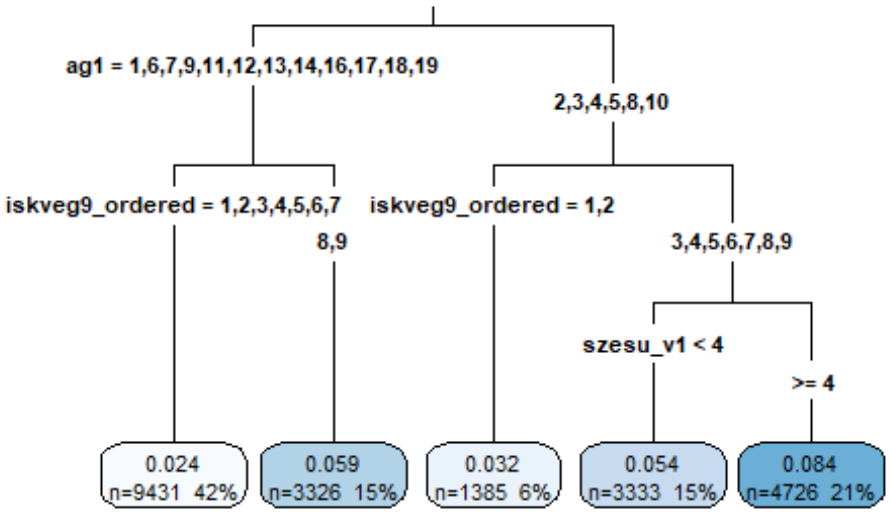


Figure 3.8 CART for 2015

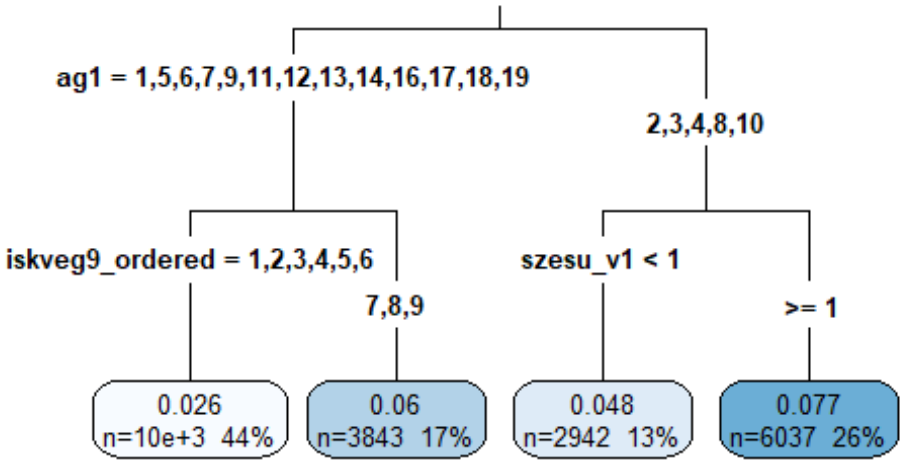


Figure 3.9 CART for 2016