Doctoral School of
Economics, Business
and Informatics

# COLLECTION OF THESIS

**Olga Takács**

## Estimation of wage functions in Hungary with machine learning algorithms

Ph.D. dissertation

**Supervisor:**

**János Vincze, DSc**
professor

Budapest, 2022

**Institute of Economics**

# COLLECTION OF THESIS

**Olga Takács**

## Estimation of wage functions in Hungary with machine learning algorithms

Ph.D. dissertation

**Supervisor:**

**János Vincze, DSc**
professor

# Contents

# 1 Previous researches and motivation

One of the most researched topics in labor economics is what factors determine wages. Question like how much education matters may arise. How much does the salary of employees increase with age? Which firms pay better: the domestic or the foreign-owned ones? In which sector do the firms pay the most and the least?

The cornerstone of the literature is the works of Mincer (1958) and Becker (1964). The human capital model they defined focused on the effect of education, age and experience on wages. In later researches new personal, company and labor market characteristics were included. (Wachtel–Betsey (1972), Blanchflower et al. (1992)) Moreover other researches sought to improve the fit of the models. (Heckman (1979), Chamberlain (1991))

Side by side with these researches, starting with Becker (1971), economists were increasingly concerned with the topic of discrimination in the labor market. The cornerstone of labor market discrimination literature is the work of Blinder (1973) and Oaxaca (1973). The Blinder-Oaxaca decomposition has since become a fundamental tool for gauging differences between groups (Jann (2008), Fortin et al. (2011)) and is often used for analyzing gender pay gap. (Weichselbaumer–Winter-Ebmer (2005)) The decomposition is to separate the explained and unexplained part within the average difference. The decomposition requires the estimations of wage functions for the examined groups.

Methodological changes were also made in the determination of wage functions and in Blinder-Oaxaca decomposition, which contributed to better understand the structural relations between wages/wage gaps and labor market characteristics. However, in the majority of literature so far, parametric estimations – usually regressions – have been used

to determine the wage functions. These estimates always require preliminary assumptions about the data generating process, which usually derive from theoretical models, but are not always properly reasoned.

In the determination of wage functions, there may still be hitherto treated as fact, but questionable or unexplained theoretical relations between variables. The inclusion or omission of these relations in the wage function estimations can mislead the thinking. In my dissertation, the use of machine learning algorithms is an innovation in the wage function estimations. Machine learning algorithms use different assumptions than traditional statistical methods do. Machine learning algorithms try to use a wide range of possible function forms, which means they can extract as much information as possible from the data in a "data-driven" way. These algorithms are able to capture the relations inherent in the data, so there is no need to make prior assumptions about the data generating process.

In the thesis, the use of machine learning algorithms is twofold. One one hand I estimate the Hungarian wage function with them to reveal the connections between variables. Some of the connections revealed by the algorithm may be ones that were already known or they may raise new questions. On the other hand, the goal of machine learning algorithms is to make the most accurate forecasts, which I use in the analysis of gender pay gap. Gender pay gap is often analysed with Blinder-Oaxaca decomposition, which requires wage function estimations for men and women separately. In my dissertation, I estimate these functions with a machine learning algorithm.

The dissertation is organized as follows. Chapter 2 summarizes the related literature on Hungary and includes the description of the database that I used in my own researches. Chapter 3 presents the machine learning algorithms and some interpretation procedures.

Chapter 4 is the edited verion of Takács–Vincze (2018), in which the Hungarian wage function was estimated with decision trees. In the researches presented in Chapter 2, the wage functions were estimated with traditional econometric methods. As a result of which the form of the function was limited, so only the relations defined in advance by the authors were included. The decision trees make it possible to estimate wages without prior assumptions about the data generation process and to examine the relations between variables in a different way. The research can also be considered as a preliminary data analysis, in which the goal is to reveal hitherto hidden relations with machine learning algorithms.

Chapter 5 and Chapter 6 relate to the gender pay gap literature. Chapter 5 is the edited version of Takács (2021). This chapter includes the description of Blinder-Oaxaca decomposition and its rewritten version for random forests. In this chapter, I compare wage functions estimated by OLS regression and generalized random forest. I compare the results of Blinder-Oaxaca decomposition estimated by these methods. My goal is to show how the choice of methodology affects the results obtained from the decomposition.

Chapter 6 is the edited version of Takács–Vincze (2020) in which I analyze the results of Blinder-Oaxaca decomposition estimated by the generalized random forest in time. I present how the explained and unexplained part obtained from the Blinder-Oaxaca decomposition change over time. At the end of the chapter, I separate the groups that contributed the most and the least to the formation and the subsistence of the average gender pay gap. Chapter 7 summarizes the results of the thesis.

# 2 Applied methods

## Database

In the dissertation, I use data from the Wage and Earnings Survey (hereinafter referred to WES) collected by the National Employment Office of Hungary. The database provides information on companies (number of employees, ownership, sector, location) and employees (education, age) along with income data (basic salary, bonuses, allowances). The dependent variable is the logarithm of average monthly earnings in each chapter. The average monthly earning consists of the basic salary and other allowances/bonuses.

I examine only full-time employees in the private sector. Furthermore, I exclude companies with fewer than 20 employees, as Lovász (2013) also did, referring to the results of Elek et al. (2009). According to Elek et al. (2009) black and gray employment is more common in smaller companies. For this reason, these companies' wage data is not reliable.

In Chapters 4 and Chapter 5, I focus on data from 2016. And in Chapter 6, I rely on the data between 2008 and 2016 when examining the evolution of gender wage gap over time. In the researches presented in Chapter 4, 5 and 6, the used explanatory variables are different: Table 1 summarizes the explanatory variables to be used in each chapter.

Table 1: Explanatory variables in researches

| Variable | Measurement | Chapter | | |
|---|---|---|---|---|
| | | 4 | 5 | 6 |
| Gender | 0: female, 1: male | X | | |
| Education | 9 categories | X | X | X |
| Age | Years | X | X | X |
| Tenure | Month | X | X | X |
| Occupation | 1-digit Hungarian ISCO | X | | |
| Occupation | 2-digit Hungarian ISCO | | X | X |
| Foreign ownership | 4 categories | X | X | X |
| State ownership | 4 categories | X | X | X |
| Size | Number of employees | X | X | X |
| Region | NUTS 2 categories | X | X | X |
| Settlement | 1: Bp., 2: city, 3: other | X | X | X |
| Sector | 1-digit NACE | X | X | X |
| Collective labor agreement on firm level | 0: no, 1: yes | X | X | |
| Collective labor agreement on sectoral level | 0: no, 1: yes | X | X | X |
| Collective labor agreement within more firms | 0: no, 1: yes | X | X | |

Source: WES

In Table 1, tenure covers the months spent at the current employer. Due to the small number of observations, I exclude sector O, which is public administration, defense and compulsory social insurance. In Chapter 6, collective agreements are merged into one variable: if a person has any type of collective agreement, the variable is 1, 0 otherwise.

## Machine learning algorithms

Among the machine learning algorithms, I use decision trees. In the case of trees, the dataset is displayed by the root which can be found at the top of the tree. At each step, the tree divides the population into two disjoint sets, to which it averages the result variable. These averages assigned to the groups represent the forecast for the group members. The dataset, the root, becomes a node and the first cut point is determined.

To determine the next cut, the algorithm examines the two leaves independently and divides one leaf into two more groups according to

one variable – the cut variable. That leaf becomes a node and two new leaves are created. Thereby the procedure creates new leaves with new cuts, it grows the decision tree. The tree stops growing when it reaches a predefined exit criterion.

In the case of algorithms using statistical tests, which I also use in my thesis, the selection of the cut variable is based on a statistical test measuring the association between the explanatory variables and the result variable. (Jiao et al. (2020)) The use of statistical tests helps to avoid the problem of overfitting, as well as the disadvantage of CART-based procedures that these algorithms favor variables with multiple potential cut points.

The disadvantage of decision trees is that their prediction error is high and they are not robust to the changes in the dataset. However, the robustness and the accuracy of predictions can be improved by aggregating results from multiple trees. Three procedures are used for this purpose: bagging, boosting and random forests. (James et al. (2013)) Random forest builds decision trees on bootstrap samples. In the cut points, the algorithm considers some of the variables as potential cut variables and not all of them. (Hastie et al. (2009))

In Chapter 4, I build a decision tree to examine the relations between wages and the explanatory variables. Among the available algorithms, I chose Conditional Inference Tree (CTree). I use random forest for those estimates where the accuracy of the forecasts plays an important role. For this reason, in Chapters 5 and 6, I estimate the wage functions for Blinder-Oaxaca decomposition with regression forests developed by Athey et al. (2019). In Chapter 4, I check the robustness of CTree with the Conditional Random Forest (CForest) which aggregates CTrees.

### Conditional Inference Tree – CTree

In the first step CTree tests the null hypothesis about the joint independence of all explanatory variables and the outcome variable. If this global null hypothesis is rejected by the algorithm, it assigns to each variable the p-value of the test statistic for the association with the result variable. The algorithm finds the explanatory variables where the null hypothesis regarding the independence of that variable and the result one can be rejected. The algorithm selects the variable with the smallest p-value which indicates that this variable has the strongest connection with the outcome variable.

In the next step, the algorithm examines all possible cut points for the selected variable. A statistical test is also used to determine which cut point is to be chosen. CTree continues cuts until a predefined exit criterion is met.

### Conditional Random Forest – CForest

CForest aggregates trees created by CTree. This forest builds trees on samples without replacement. Moreove, the predictions are given as the average of the observations with which the observation to be predicted falls on the same leaf in each tree. (Levshina (2020))

### Conditional variable importance

Conditional variable importance expresses how much the input variables affect the outcome variable on a quantitative scale. (Inglis et al. (2022)) According to this, it is easier to decide which variables are more and which are less important from the point of view of the outcome variable, without having to examine the entire tree in detail. In my dissertation, I use a variable importance measure based on conditional

permutation procedure developed by Strobl et al. (2008).

## Generalized Random Forest – Regression Forest

Generalized Random Forest was created by Athey et al. (2019). This algorithm uses some features of the random forest, such as sequential binary cuts, sampling and limiting the number of possible cut variables at the cut points. However, the algorithm also takes into consideration another essential property, the honesty: the algorithm divides the learning sample into two parts and performs the cuts on one half, and it quantifies the estimates for the leaves on the other half. For regression forests, sample means sampling without replacement.

Making predictions in regression forest is to average the observations that fall on the same leaf with the observation to be predicted. Thus, an observation that is close to the observation to be predicted in the variable space has higher weight in the average.

### Classification Analysis– CLAN

The Classification Analysis (CLAN) was developed by Chernozhukov et al. (2018). The goal of the authors was to develop a procedure that can be used to examine the main properties of conditional average treatment effects calculated with machine learning algorithms. The conditional average treatment effect shows the average difference between the treated and the untreated group in randomized experiments.

CLAN seeks to answer the question of what characteristics the most and the least affected groups possess. "The most" affected group includes the largest, and "the least" affected group includes the observations with the smallest treatment effect. The vectors describing the characteristics of these groups show the general characteristics of the members of the two groups. The elements of the vectors are calcu-

lated directly from the observed variables. Furthermore, the difference of these vectors can be calculated as well, which shows the average difference between the two groups.

**Random forest for gender pay gap**

In Chapters 5 and 6, I approximate the conditional expected value functions for wages with OLS regression and regression forest. For OLS estimation, the Blinder-Oaxaca decomposition can be written as:

$$\overline{Y_M} - \overline{Y_F} = \left(\overline{X_M} - \overline{X_F}\right)\beta_M - \overline{X_F}\left(\beta_M - \beta_F\right) + \text{bias}, \qquad (1)$$

where $\overline{Y_M} - \overline{Y_F}$ captures the difference between the average wage of male ($M$) and female ($F$) subsets which is the raw pay gap. $\overline{X_M}$ and $\overline{X_F}$ are the average values of the explanatory variables for men and women, respectively. $\beta_M$ and $\beta_F$ are the coefficients of the OLS regressions estimated separately for the two genders. On the right side of (1), the first part captures the composition effect, the second is the wage structure effect. Composition effect indicates the part of the raw gap which can be explained by the labour market characteristics, wage structure effect is the part of the decomposition which cannot be explained by them. In the case of OLS, the bias on the training sample is zero by definition, while it may differ from zero on the test sample.

The form of (1) is the same as the original one in Blinder (1973) and Oaxaca (1973). The authors believed that the wage structure effect perfectly measures discrimination. However, subsequent studies have shown that this part also contains other statistical errors: for example, the effect of omitted variables or the incorrect measurement of work experience. (Reilly (2001)) For this reason, the wage structure effect provides important information regarding gender differences, but it is worth being careful when interpreting it.

The heterogeneous wage structure effect for the $i$th observation, based on the characteristics of the person observed, can be written in the following form:

$$HWS(i) = X_i^{'}(\beta_M - \beta_F), \tag{2}$$

where $X_i$ is the vector of the observed characteristics of the $i$th person, while $\beta_M$ and $\beta_F$ are the coefficients from the OLS regression estimations on the male and female subsamples, respectively. The average of the heterogeneous wage structure effects calculated for women is the same as the second term on the right side of (1), the average wage structure effect.

In the case of the regression forest, $\beta$ coefficients cannot be estimated, only the predictions directly. Thus, the following equation shows the extension of Blinder-Oaxaca decomposition to random forest:

$$\overline{Y_M} - \overline{Y_F} = \left(\overline{P_M} - \overline{P_{F(M)}}\right) - \left(\overline{P_{F(M)}} - \overline{P_F}\right) + \text{bias}, \tag{3}$$

where $\overline{P_M}$ is the average of the predictions made by the random forest trained on the male subset, while $\overline{P_F}$ is the average of the predictions made by the forest trained on the female subset. $\overline{P_{F(M)}}$ is the average of the counterfactual predictions which means that the average of the predictions on female subset made by the random forest trained on the male subset.

Heterogeneous wage structure effect for the $i$th person with the wage functions estimated by random forests is:

$$HWS(i) = P_M(i) - P_F(i), \tag{4}$$

where $P_M(i)$ and $P_F(i)$ are the predicted wages estimated for $i$ by the random forest trained on the male and the female subset, respectively.

I measure the impact of the explanatory variables on the wage structure effect by analyzing the heterogeneous wage structure effects by variable. In the case of OLS and the forest, the heterogeneous wage structure effect can be calculated for women only. For this reason, I perform the analysis of heterogeneous wage structure effect only for women. If the heterogeneous wage structure effect is positive, then the woman is under-; and if it is negative, she is overpriced compared to a man having the same labour market characteristics.

In the first step I compare the averages and the distributions of the wage structure effects calculated by the two methods. In the second step, I compare the averages calculated from the heterogeneous wage structure effects for the categories of the explanatory variables with descriptive statistical methods in order to see how the variables affect the heterogeneous wage structure effect.

In Chapter 6, I use Classification Analysis, CLAN to examine the most extreme groups. I analyze the persons belonging to the top and bottom ten percent of the distribution of heterogeneous wage structure effect. CLAN shows the general characteristics of the observations belonging to these groups, and their analysis shows the characteristics of the women in Hungarian labour market who are priced the least or the most differently than men with the same characteristics.

# 3 Results of the thesis

## Estimation of wage function with Conditional Inference Trees

- The trees and the variable importance table also clearly showed that occupation is the most important explanatory variable, which is in agreement with the literature. As occupations become more

and more routine and the FEOR codes take higher and higher values, the earnings decrease.

- The second most important variable is education, for which the order of the variable categories has been seen in the trees: higher degree means higher wage. However, the income-increasing effect of education is not uniform: in the case of tertiary education, the wage level jumps.

- According to the optimal tree, sector and foreign ownership are equally important, but in the medium tree the effect of foreign ownership was even more significant. The importance of the sector increased if surrogates were allowed and the analysis of the surrogates showed that this variable can be a good substitute for all other variables. These results confirm the findings of Kőrösi (2006) that sector is more significant than foreign ownership and that two are strongly correlated.

- Foreign ownership strongly affects wages. Companies with any degree of foreign ownership pay higher wages than those with zero percent of foreign ownership. This result verifies the same finding of Kőrösi (2006). The order of foreign ownership can be seen in the tree, the relationship between the two variables is monotonic.

- According to the variable importance measures, age is not nearly as important as expected. The analysis of the trees indicated a non-monotonic relationship between wages and age, but it is not necessarily squared as shown by Gábor (2008). The length of service spent at the current employer proved to be more important than age. This result allows me to conclude that the specific knowledge acquired at the given workplace may be more decisive

in the terms of wages, as opposed to the general work experience, which was proxied by age.

- Regarding the regions, Central Hungary and Central Transdanubia are the areas where employees earn more, which is in line with the results of Fazekas (2005) and Bartus (2003). Fazekas (2005) pointed out that after the regime change the location of foreign companies was influenced by the distance to the western border, while the location of the Hungarian companies was not affected by this factor. However, when examining the surrogates, it appeared that region was a good surrogate for the foreign and the state ownership variables. So, my own results partially confirm the statement of Fazekas (2005) and show that geographic factors affect the location of both types of companies, but this factor is not necessarily the distance to the western border.

- The capital city (Budapest) usually belongs to the higher paying areas. However, in many cases, the other type of settlement was at the same node as Budapest in the optimal tree. For this reason, the result of Köllő (2003), according to which Budapest has a significant wage advantage compared to other settlement types, cannot be unequivocally justified.

- Gender discrimination is visible in CTree, as men are always in the higher paying groups with one exception. Furthermore, confirming the results of Csillag (2006), the occupational division of men and women is less and less present in Hungarian labor market, however, the division of genders by sectors is visible.

- Rigó (2012) showed that the coverage of collective agreements in Hungary is not independent from the sectors and the geographical regions. Accordint to my own results, the sector and the region can be a good substitute for collective agreements. Furthermore,

Rigó (2012) argued that the impact of collective agreements on wages is low, which was supported by the variable importance measures. CTree was able to verify both of Rigó (2012)'s findings.

## Gender pay gap in Hungary: the comparison of the results of an OLS regression and a regression forest

- A comparison of the estimates obtained from the OLS regression and the regression forest shows that the wage forecasts made by the forests are more accurate. The use of random forests is justified by the fact that the bias appearing in Blinder-Oaxaca decomposition is acceptably small. So the change in methodology does not significantly affect the bias and the wage forecasts are more accurate.

- According to the Blinder-Oaxaca decomposition, composition effect is lower in the case of the random forest, which means that the wage structure effect is higher. However, the average wage structure effects calculated by the two methods are close to one another.

- According to the analysis of the heterogeneous wage structure effects, although the averages are similar, the distributions differ significantly and the correlation between the estimates is only moderate. The results show that, based on the two methods, there is a difference about how much a woman is mispriced compared to a man having the same characteristics. According to the analysis of heterogeneous wage structure effects, if several observations belong to a category, the average wage structure effects calculated by the two methods are closer to each other. The descriptive statistical analysis indicated that there is no one variable that is responsible for the difference in the average wage structure

effects calculated with the two methods. The difference between the groups with fewer observations can attribute to the difference in the distributions and the moderate correlation between the estimation results.

- My results confirm the findings of Weichselbaumer–Winter-Ebmer (2005), according to which different methods lead to similar results for decomposition, but the variables are evaluated differently by each method. So if we want to examine the effect of the variables, the methodology must be chosen carefully, because it can affect the size of the composition and the wage structure effects measured for the variables.

## Change of the heterogeneous wage structure effect over time and the most extreme groups

- According to the comparison of the performance of the regression forest and the OLS regression including interactions, the forest has a more accurate prediction performance than the OLS estimation. The mean squared errors calculated with both methods on the test sample increase, but the errors of the regression forest are still lower.

- When comparing the parts of Blinder-Oaxaca decomposition, the composition effects obtained with the regression forest are generally lower than those obtained with OLS estimation, but they have the same magnitude and move in the same directions. The composition effect was initially negative and showed an increasing trend. The raw wage gap increased over time, the wage structure effect can be considered constant during the period, which is in agreement with Lovász (2009)'s results. The author showed that after the change of regime in Hungary, the wage structure effect

decreased drastically, and from 1992 it moderated at a much lower pace and almost stagnated.

- The negative composition effect can be a long-term result of the processes following the change of regime. After it, the return on education increased in Hungary, and since women were more educated than men on average, this factor induced the growth of the composition effect. (Kertesi–Köllő (1996), Brainerd (2000)) In addition, the labor supply for women decreased in Hungary during this period (Brainerd (2000)), presumably because lower-educated, less productive women left the labor market. Thus, the overall labor market characteristics of women have improved.

- On Hungarian labor market, a significant part of men have a secondary education degree without high school diploma, while the majority of women have high school diploma. The trends between the years are similar for both genders: the proportion of people with higher education is increasing, while the proportion of people without high school diploma is decreasing. However, this transition is happening very slowly.

- Between 2008 and 2016, the proportion of men working in 100 percent foreign-owned companies increased, while the proportion of women decreased. This transition – since foreign companies generally pay higher wages (Kőrösi (2006)) – has influenced the growth of the composition effect.

- The analysis of the heterogeneous wage structure effect showed that the relations between average effects and education levels is not monotonic. The average wage structure effects initially increase more with age and the length of service, then decrease slightly, but are still positive, which confirms the results of Gábor (2008).

- When comparing the average wage structure effects and the foreign share, a monotonic relationship can be seen: as foreign ownership decreases, the average effect also decreases. Köllő (2000) argues the possibility that after the change of regime, those companies where the gender wage gap were larger were taken over by foreigners, as a result of which the average wage gap increased at foreign companies, while it decreased at domestic ones. Presumably, the long-term effect of this change can also be observed in my own results.

- According to CLAN, the least effected women, for whom the wage structure effect was low, work for not very large domestic companies in the service sector in Central Hungary. These women may be paid the same or even better than men. The women with higher wage structure effects usually work for larger, foreign-owned manufacturing companies in Transdanubia.

- The results of the CLAN support the monopsony labor market theory. According to the theory, due to the different commuting and search costs between the genders, as well as the importance of non-financial benefits, the wage elasticity of labour supply to a firm differs for men and women, which can provide market power for employers over their workers. (Hirsch (2009), Hirsch–Schnabel (2010), Webber (2013), Heinze–Wolf (2010)) In the central region, women can get new jobs more easily, they have a better position on the labor market, the competition among employers is presumably greater, which is why the companies' monopsony power may be less in this geographic area. In addition, in the least effected group, there is a higher proportion of those who have at least a high school diploma – high school graduates and tertiary graduates together – which strengthens further their position on the labor market. Of course, even in this segment, it may happen

that the elasticity of labor supply to a firm differs by gender but the chance that the gender wage gap is due to the monopsony power of the company is already lower. Differences in productivity and/or discrimination between genders may still appear.

# 4 References


Athey, S. – Tibshirani, J. – Wager, S. (2019): Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, DOI: http://dx.doi.org/doi10.1214/18-AOS1709.

Bartus, T. (2003): Ingázás. In Fazekas, K., eds., *Munkaerőpiaci tükör 2003*, pp. 88–101. MTA Közgazdaságtudományi Kutatóközpont.

Becker, G. S. (1964): *Human Capital*. Columbia University Press for the National Bureau of Economic Research, New York, NY, first edition.

Becker, G. S. (1971): *The Economics of Discrimination*. The University of Chicago Press, Chicago.

Blanchflower, D. G. – Oswald, A. J. – Sanfey, P. (1992): Wages, profits and rent-sharing. Working Paper 4222, National Bureau of Economic Research, DOI: http://dx.doi.org/doi10.3386/w4222.

Blinder, A. S. (1973): Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8(4):436–455, DOI: http://dx.doi.org/doi10.2307/144855.

Brainerd, E. (2000): Women in transition: Changes in gender wage differentials in eastern europe and the former soviet union. *Industrial and Labor Relations Review*, 54(1):138–162, DOI: http://dx.doi.org/doi10.2307/2696036.

Chamberlain, G. (1991): Quantile regression, censoring, and the structure of wages. In Sims, C. A., eds., *Advances in Econometrics*, pp. 171–210. Cambridge University Press, DOI: http://dx.doi.org/doi10.1017/ccol0521444594.005.

Chernozhukov, V. – Demirer, M. – Duflo, E. – Fernández-Val, I. (2018): Generic machine learning inference on heterogenous treatment effects in randomized experiments. NBER Working Paper 24678, National Bureau of Econimic Research, DOI: http://dx.doi.org/doi10.3386/w24678.

Csillag, M. (2006): Női munka és nemek szerinti kereseti különbségek a késő szocializmustól napjainkig. In Fazekas, K. – Kézdi, G., eds., *Munkaerőpiaci tükör 2006*, pp. 100–105. MTA Közgazdaságtudományi Intézet.

Elek, P. – Scharle, Á. – Szabó, B. – Szabó, P. A. (2009): A bérekhez kapcsolódó adóeltitkolás magyarországon. Közpénzügyi füzetek 23, Pénzügyminisztérium.

Fazekas, K. (2005): A hazai és a külföldi tulajdonú vállalkozások területi koncentrációjának hatása a foglalkoztatás és munkanélküliség területi különbségeire. In Faluvégi, A. – Fazekas, K. – Nemes-Nagy, J. – Németh, N., eds., *A hely és a fej: Munakpiac és regionalitás Magyarországon, Budapest*, pp. 47–74. MTA Közgazdaságtudományi Kutatóközpont.

Fortin, N. – Lemieux, T. – Firpo, S. (2011): Decomposition Methods in Economics. In Ashenfelter, O. – Card, D., eds., *Handbook of Labor Economics*, volume 4 of *Handbook of Labor Economics*, chapter 1, pp. 1–102. Elsevier, DOI: http://dx.doi.org/doi10.1016/s0169-7218(11)00407-2.

Gábor, R. I. (2008): A hiányzó láncszem? Életpálya-keresetek és keresetingadozás. *Közgazdasági Szemle*, 55(december):1057–1074.

Hastie, T. – Tibshirani, R. – Friedman, J. (2009): *The elements of statistical learning: data mining, inference and prediction*. Springer, 2nd edition, DOI: http://dx.doi.org/doi10.1007/b94608.

Heckman, J. J. (1979): Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, DOI: http://dx.doi.org/doi10.2307/1912352.

Heinze, A. – Wolf, E. (2010): The intra-firm gender wage gap: a new view on wage differentials based on linked employer–employee data. *Journal of Population Economics*, 23(3):851–879, DOI: http://dx.doi.org/doi10.1007/s00148-008-0229-0.

Hirsch, B. (2009): The gender pay gap under duopsony: Joan robinson meets harold hotelling. *Scottish Journal of Political Economy*, 56(5):543–558, DOI: http://dx.doi.org/doi10.1111/j.1467-9485.2009.00497.x.

Hirsch, Borisand Schank, T. – Schnabel, C. (2010): Differences in labor supply to monopsonistic firms and the gender pay gap: An empirical analysis using linked employer-employee data from germany. *Journal of Labor Economics*, 28(2):291–330, DOI: http://dx.doi.org/doi10.1086/651208.

Inglis, A. – Parnell, A. – Hurley, C. B. (2022): Visualizing variable importance and variable interaction effects in machine learning models. *Journal of Computational and Graphical Statistics*, pp. 1–13, DOI: http://dx.doi.org/doi10.1080/10618600.2021.2007935.

James, G. – Witten, D. – Hastie, T. – Tibshirani, R. (2013): *An Introduction to Statistical Learning: with Applications in R*. Springer, DOI: http://dx.doi.org/doi10.1007/978-1-4614-7138-7.

Jann, B. (2008): The blinder–oaxaca decomposition for linear regression models. *Stata Journal: Promoting communications on statistics and Stata*, 8(4):453–479, DOI: http://dx.doi.org/doi10.1177/1536867x0800800401.

Jiao, S. R. – Song, J. – Liu, B. (2020): A review of decision tree classification algorithms for continuous variables. *Journal of Physics: Conference Series*, 1651(1):012–083, DOI: http://dx.doi.org/doi10.1088/1742-6596/1651/1/012083.

Kertesi, G. – Köllő, J. (1996): A bér alakulását meghatározó tényezők. In Halpern, L., eds., *Bérköltség és versenyképesség*, pp. 74–76. MTA Közgazdaságtudományi Intézet, január.

Köllő, J. (2000): Tulajdoni szektorok. In Fazekas, K., eds., *Munkaerőpiaci tükör 2000*, pp. 100–106. MTA Közgazdaságtudományi Kutatóközpont.

Köllő, J. (2003): Regionális kereseti és bérköltség-különbségek. In Fazekas, K., eds., *Munkaerőpiaci tükör 2003*, pp. 65–78. MTA Közgazdaságtudományi Kutatóközpont.

Kőrösi, G. (2006): Vállalatok közti bérkülönbségek dinamikája. In Fazekas, K. – Kézdi, G., eds., *Munkaerőpiaci tükör 2006*, pp. 48–59. MTA Közgazdaságtudományi Kutatóközpont.

Levshina, N. (2020): Conditional inference trees and random forests. In Paquot, M. – Gries, S. T., eds., *A Practical Handbook of Corpus Linguistics*, pp. 611–643. Springer International Publishing, Cham, DOI: http://dx.doi.org/doi10.1007/978-3-030-46216-1_25.

Lovász, A. (2009): A verseny hatása a női-férfi bérkülönbségre magyarországon 1986 és 2003 között. In Fazekas, K. – Lovász, A. – Telegdy, Á., eds., *Munkaerőpiaci tükör 2009*, pp. 149–158. MTA Közgazdaságtudományi Intézet.

Lovász, A. (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*, 60:1057–1074.

Mincer, J. (1958): Investment in human capital and personal income distribution. *Journal of Political Economy*, 66(4):281–302, DOI: http://dx.doi.org/doi10.1086/258055.

Oaxaca, R. (1973): Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709, DOI: http://dx.doi.org/doi10.2307/2525981.

Reilly, B. (2001): A nemek közötti bérkülönbségek elemzésének statisztikai módszerei. *Statisztikai Szemle*, 79(1):5–17.

Rigó, M. (2012): A szakszervezeti bérrés becslése magyarországon. In Fazekas, K. – Benczúr, P. – Telegdy, Á., eds., *Munkaerőpiaci tükör 2012*, pp. 200–214. MTA Közgazdaságtudományi Kutatóközpont.

Strobl, C. – Boulesteix, A.-L. – Kneib, T. – Augustin, T. – Zeileis, A. (2008): Conditional variable importance for random forests. *BMC bioinformatics*, 9:307, DOI: http://dx.doi.org/doi10.1186/1471-2105-9-307.

Takács, O. (2021): Nemek közötti bérkülönbségek magyarországon: a véletlenerdő- és az ols-becslésen alapuló blinder–oaxaca-dekompozíció eredményeinek összehasonlítása. *Statisztikai Szemle*, 99(1):5–45, DOI: http://dx.doi.org/doi10.20311/stat2021.1.hu0005.

Takács, O. – Vincze, J. (2018): Bérelőrejelzések – prediktorok és tanulságok. *Közgazdasági Szemle*, 55(6):592–618, DOI: http://dx.doi.org/doi10.18414/ksz.2018.6.592.

Takács, O. – Vincze, J. (2020): The gender-dependent structure of wages in hungary: results using machine learning techniques. KRTK-

KTI műhelytanulmány 44, MTA Közgazdaság- és Regionális Tudományi Kutatóközpont.

Wachtel, H. M. – Betsey, C. (1972): Employment at low wages. *The Review of Economics and Statistics*, 54(2):121–129, DOI: http://dx.doi.org/doi10.2307/1926272.

Webber, D. A. (2013): Firm-level monopsony and the gender pay gap. IZA Discussion Papers 7343, Institute of Labor Economics (IZA).

Weichselbaumer, D. – Winter-Ebmer, R. (2005): A meta-analysis of the international gender wage gap. *Journal of Economic Surveys*, 19(3):479–511, DOI: http://dx.doi.org/doi10.1111/j.0950-0804.2005.00256.x.

# 5 Publications in the topic of the thesis by the candidate

Takács, O. – Vincze, J. (2018a): Bérelőrejelzések - prediktorok és tanulságok. *Közgazdasági Szemle*, 55(6):592–618, DOI: http://dx.doi.org/10.18414/ksz.2018.6.592.

Takács, O. – Vincze, J. (2018b): The within-job gender pay gap in Hungary. Közgazdaságtudományi Intézet műhelytanulmány 34, MTA Közgazdaság- és Regionális Tudományi Kutatóközpont.

Takács, O. – Vincze, J. (2019a): Blinder-oaxaca decomposition with recursive tree-based methods: a technical note. Közgazdaságtudományi Intézet műhelytanulmány 23, MTA Közgazdaság- és Regionális Tudományi Kutatóközpont.

Takács, O. – Vincze, J. (2019b): The gender pay gap in hungary: new results with a new methodology. Közgazdaságtudományi Intézet műhelytanulmány 24, MTA Közgazdaság- és Regionális Tudományi Kutatóközpont.

Takács, O. – Vincze, J. (2020): The gender-dependent structure of wages in Hungary: results using machine learning techniques. CERS-IE WORKING PAPERS 2044, Institute of Economics, Centre for Economic and Regional Studies.

Takács, O. (2021): Nemek közötti bér különbségek Magyarországon: a véletlenerdő- és az OLS-becslésen alapuló Blinder–Oaxaca dekompozíció eredményeinek összehasonlítása. *Statisztikai Szemle*, 99(1): 5–45. DOI: http://dx.doi.org/10.20311/stat2021.1.hu0005