INSTITUTE OF ECONOMICS
Centre for Economic and Regional Studies

KRTK KTI

Közgazdaság- és Regionális Tudományi Kutatóközpont
KÖZGAZDASÁG-TUDOMÁNYI INTÉZET

MŰHELYTANULMÁNYOK DISCUSSION PAPERS

MT-DP – 2019/23

# Blinder-Oaxaca decomposition with recursive tree-based methods: a technical note

OLGA TAKÁCS – JÁNOS VINCZE

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

Discussion papers
MT-DP – 2019/23

Institute of Economics, Centre for Economic and Regional Studies,

# Blinder-Oaxaca decomposition with recursive tree-based methods: a technical note

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

# Blinder-Oaxaca decomposition with recursive tree-based methods: a technical note

Olga Takács – János Vincze

## Abstract

The Blinder-Oaxaca decomposition was developed in order to detect and characterize discriminatory treatment, and one of its most frequent use has been the study of wage discrimination. It recognizes that the mere difference between the average wages of two groups may not mean discrimination (in a very wide sense of the word), but the difference can be due to different characteristics the groups possess. It decomposes average differences in the variable of interest into two parts: one explained by observable features of the two group, and an unexplained part, which may signal discrimination. The methodology was originally developed for OLS estimates, but it has been generalized in several nonlinear directions. In this paper we describe a further extension of the basic idea: we apply Random Forest (RF) regression to estimate the explained and unexplained parts, and then we employ the CART (Classification and Regression Tree) methodology to identify the groups for which discrimination is most or least severe.

JEL codes: C10, C14, C18

Keywords: Oaxaca-Blinder decomposition, Random Forest Regression. CART

# A foglalkozáson belüli nemek közötti bérkülönbség Magyarországon

Takács Olga – Vincze János

## Összefoglaló

A Blinder–Oaxaca-dekompozíciót a diszkrimináció statisztikai vizsgálatára fejlesztették ki. Ennek egyik legfontosabb felhasználási területe a bérdiszkrimináció elemzése. Az módszer szerint pusztán két társadalmi csoport átlagbéreinek különbsége nem jelent diszkriminációt, hiszen lehetnek olyan releváns különbségek a két csoport tagjainak tulajdonságai között, amelyek a bérkülönbséget okozzák. A dekompozíció két részre bontja a bérkülönbséget: a csoportok megfigyelhető tulajdonságaival magyarázott, valamint egy nem magyarázott részre, ami potenciálisan a bérdiszkrimináció. (Természetesen a tulajdonságok különbségei is származhatnak – például az oktatási rendszerből eredő – diszkriminációból, de az elemzésben csak a bérezés terén jelentkező diszkrimináció a kérdés.) Az eredetileg OLS-kontextusban kifejlesztett módszertant ebben a tanulmányban kiterjesztjük olyan módon, hogy a magyarázott és nem magyarázott részek becslésére véletlen erdő regressziót használunk, majd a klasszifikációs és regressziós fa technika segítségével azonosítjuk azokat a csoportokat, amelyek leginkább vagy legkevésbé vannak kitéve diszkriminációnak.

JEL: C10, C14, C18

Tárgyszavak: Blinder-Oaxaca dekompozíció, véletlen erdő regresszió, klasszifikációs és regressziós fa

# Blinder-Oaxaca decomposition with recursive tree-based methods: a technical note

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

The Blinder-Oaxaca decomposition was developed in order to detect and characterize discriminatory treatment, and one of its most frequent use has been the study of wage discrimination. It recognizes that the mere difference between the average wages of two groups may not mean discrimination (in a very wide sense of the word), but the difference can be due to different characteristics the groups possess. It decomposes average differences in the variable of interest into two parts: one explained by observable features of the two group, and an unexplained part, which may signal discrimination. The methodology was originally developed for OLS estimates, but it has been generalized in several nonlinear directions. In this paper we describe a further extension of the basic idea: we apply Random Forest (RF) regression to estimate the explained and unexplained parts, and then we employ the CART (Classification and Regression Tree) methodology to identify the groups for which discrimination is most or least severe.

JEL codes: C10, C14, C18

Keywords: Oaxaca-Blinder decomposition, Random Forest Regression. CART

## 1 Introduction

Traditional statistical methods are of limited utility when there are many possible "predictor" variables, and when complex interactions exist in the data. To overcome these problems researchers have turned to "machine learning" algorithms that automatize variable and functional form selection. Experience in several fields have shown that these methods perform better as predictors for problems characterized by the above features.

Tree-based methods make up one group of such algorithms that have gained currency in many applications in recent years. Below we propose a combination of two of them (CART and RF) for generalizing the traditional regression-based methodology of the Blinder-Oaxaca decomposition. The combination intends to exploit the relative strengths of these algorithms: RF is superior as a predictor, whereas a CART's results have much better interpretability. In Section 2 we give a short overview of tree-based methods and, in Section 3, of the Blinder-Oaxaca decomposition. Section 4 details the new methodology which is illustrated by a wage discrimination example in Section 5. Section 6 gives a summary.

## 2 Tree-based methods and their properties

### *Growing a tree recursively*

To be concrete the detailed description below addresses the binary classification problem with negative entropy as a measure of the goodness of fit, but at the end we give the necessary modifications for other frameworks.

Output data define a binary distribution over the two classes they belong to. This distribution has an entropy, reflecting the uncertainty one faces when wishing to classify the objects without the knowledge of any explanatory variables. Tree building is in essence an entropy reduction process. At the beginning consider each explanatory variables (features) and calculate by how much total entropy would be reduced if one were to split the full sample in two, based on the variable in question. If a variable has many possible values then there are many (possibly infinitely many) splits, and one must choose the one with the highest reduction in entropy. After considering each variable in turn, select the one with the highest entropy reduction capacity, and perform the corresponding split of the sample. Graphically this is equivalent to forming two nodes in a tree whose parent node is the root. Geometrically a partition of the input space is the result. Entropy reduction can be viewed alternatively as purifying: the new nodes are purer than the root node, in the sense that the observations belonging to them are more homogeneous. Tree-growing is a recursive process. In the next step each descendant node is considered likewise, and new nodes are added by the same procedure. In principle this tree-growing process can lead to perfect purification (where each final node contains objects belonging to the same class), but, in practice, researchers apply some stopping criterion when, for instance, the number of objects in the final nodes should not be below a certain threshold.

For the classification problem other impurity measures can also be used, such as the Gini-measure. Trees can be grown to continuous response variables (the regression tree). In that case the most usual is to measure the goodness of fit with the mean squared error metric, but tree-growing can accommodate other measures as well. Athey-Imbens, 2016 introduced causal trees where trees are grown focussing on maximizing causal effects. There we have a binary causal variable (treatment and non-treatment cases), with the average treatment effect at each subset of the input space defined as: average of treated – average of non-treated. A causal tree cuts the tree at each node by maximizing the increase at each step in the average treatment effect.

It is clear that at the end we find a fully grown tree (if there is no stopping criterion) which gives a perfect fit, and therefore would not be very useful for prediction (an obvious case of overfitting). Still tree growing provides much information since the path to the full-grown tree is also important, it shows an optimal way to reach that. As usual overfitting leads to high variance, and it must be controlled. To make tree-growing a successful predictive device the bias-variance trade-off must be dealt with. Different approaches have been developed to use trees to get a prediction that is validated.

### *The CART (Classification and Regression Tree): pruning the tree*

The tree built in the above manner can be regarded as a non-parametric estimate of a two-valued function, where the procedure divides the input space into mutually exclusive regions, and assigns each observation to one of the classes depending on the region (leaf or final node) it belongs to. An alternative interpretation assigns a probability based on the relative frequencies of the corresponding region (final node), when the final nodes are not completely pure. There exist general theorems that assert that with a very large number of observations this estimate can be considered unbiased. However, it is also recognized that a very large (finely tuned) tree probably overfits (i.e. accommodates noise), resulting in reduced predictive abilities. Therefore, CART prunes the initially built tree using complexity cost pruning. In the first step of pruning one finds the best subtree, in the sense of least entropy or impurity, for a number of complexity classes, where a tree is more complex if it has more leaves. Then a validation procedure compares the best subtrees´ generalization capabilities by cross-validation techniques, and the one with the best predictive score is chosen as the end product of the procedure. Concrete implementations may differ in the choice of complexity cost, and in the validation procedure.

### *Using and interpreting the CART outcome*

The final tree can be interpreted as a decision tree where at each node some temporary classification decision is made, leading to final decisions concerning where to classify a certain object. This vane be taken literally, as Lewis, 2000 searches for a clinical decision rule via a CART analysis. For any new observation one has to find its region in the input space, and make the corresponding classification as a prediction. The alternative interpretation again is a probabilistic judgment, rather than a "yes-no" decision. For regression trees the prediction equals the average at each node, thus the estimator is basically a step function.

One possible use of a tree is to evaluate the relative importance of explanatory variables. Intuitively one may think that it suggests that important variables are those that have many and closer to the root splits in them. Indeed, researchers have developed formal indicators to measure the relative importance of explanatory variables, based on the entropy reduction work they do (see Ishwaran, 2007).

Another possible use of CART models is by varying the input space: we can include (suspect) variables (either deemed as relevant or irrelevant), and see how they appear in the best decision tree. We can adapt the idea of Granger-causality as well: does the inclusion of a variable significantly improve the predictive performance of the model or not? As the CART algorithm does not lead automatically to a better in-sample fit, after adding a new variable this question can (sometimes) be evaluated in a two-valued logic context, in contrast to Granger-causality where the measure of significance depends on the validity of maintained probabilistic assumptions.

Finally, CART algorithms can be applied for "audience segmentation", as they are used in public health applications. One can identify non-trivial segments of society by their homogenous behaviour, enabling policy makers to adjust interventions targeted to these different groups. This is similar to cluster analysis, but in a supervised learning context: we have a definite measure by which we judge homogeneity.

### The Random Forest (RF) algorithm

CART (see Breiman et al. (1984)) is a greedy algorithm, as it drives at each step to achieve maximal purity increase. This results in higher variance, and instability (small changes in samples lead to large changes in the tree).  Bagging is an extension that addresses this problem by growing many trees, but on bootstrap samples. Bootstrapping can be regarded as an alternative way of validation, and accordingly bagging does not use pruning, rather it averages over many large and unpruned trees.

A Random Forest Regression (see Breiman, 2001) is also constructed from a collection of Regression Trees, the number of trees is a parameter set by the researcher. The prediction (estimate) a Random Forest regression gives is the average of the constituent trees' predictions. Random Forest improves on bagging by randomizing variable choice at each cut-point, at each node only a random subset of explanatory variables are considered for a split. The cardinality of that subset is another parameter of the algorithm.

The main advantage of Random Forests is that the random and restricted manner of splitting achieves de-correlation among the many trees, while unbiasedness is not jeopardized. (Hastie et al., 2017). Varian, 2015 proposed Random Forests for econometricians by citing Howard and Bowles, 2012 who asserted that Random Forests were the most successful general-purpose predictive algorithm. Wager and Athey (2017) argue that Random Forest 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, when both the weights and the proximity are determined in a data-driven way. All in all, with the presence of significant non-linearities, and with a relative abundance of explanatory variables Random Forest seems to be a successful and well-attested predictive methodology.

Though an outstanding method for prediction Random Forest regression has a problem: the results are not easily interpretable variable-wise. The demand for assessing the separate role of variables (their individual explanatory power) led to the proposal of several variable importance measures. There exists a permutation based MSE reduction indicator, that works like this (see Grömping, 2009). As trees are grown from bootstrap samples a number of out-of-the-bag (OOB) observations belong to each tree, namely those data points that are not included in the sample for that particular tree. One can then calculate the prediction MSE on OOB data for each tree. Now the idea is that if a variable is unimportant it does not matter whether the predictions are generated with the help of their true values, or are calculated from a random permutation of the true data. (The permutation shuffles only the values of the variable in question.) Then one can calculate the difference between the true and the permuted SSE, which must be small if the variable is unimportant. By averaging all such differences over all trees one obtains a measure of variable importance. This measure is obviously ad hoc. One can use it in two ways: determining the importance ranking of variables, and by calculating relative importance shares for each variable.

## 3 The Blinder – Oaxaca decomposition

Though more generally valid for any two groups and any variable of interest, in the description below we refer to the two groups as males and females, and the variable of interest as wages. To apply the traditional Blinder-Oaxaca methodology (see Jann, 2008) one needs to run three linear regressions for the OLS-based Oaxaca-Blinder decomposition: one for the female and male subsamples, respectively, and a reference for the pooled sample. Several reference models have been proposed in the literature. Then we have the identity:

$$\bar{Y}_M - \bar{Y}_F = \bar{X}_M\beta_M - \bar{X}_F\beta_F = (\bar{X}_M - \bar{X}_F)\beta_R + \bar{X}_M(\beta_M - \beta_F) + \bar{X}_F(\beta_R - \beta_F), \tag{1}$$

where $\bar{Y}$ and $\bar{X}$ are the average of groups labelled by M (male) and F (female) here. In this equation the first part on the right-hand side is the explained part and the rest measures the unexplained part. As the sample average equals the average prediction (if a constant is included in the regression) it is indeed a decomposition of the averages, which obviously depends on the reference model. To calculate the decomposition, one must calculate the raw difference $(\bar{Y}_M - \bar{Y}_F)$ and the explained part $((\bar{X}_M - \bar{X}_M)\beta_R)$, thus the determination of $\beta_M$ and $\beta_F$ are unnecessary.

In some cases, the reference model is taken to be one of the group models, say M. Then the formula becomes simpler:

$$\bar{Y}_M - \bar{Y}_F = \bar{X}_M\beta_M - \bar{X}_F\beta_F = (\bar{X}_M - \bar{X}_F)\beta_M + \bar{X}_F(\beta_M - \beta_F).$$

As $((\bar{X}_M - \bar{X}_M)\beta_R)$ is an inner-product the explained part, in its turn, can be decomposed variable-wise. Clearly a variable's effect on the explained part is higher if it has a larger coefficient in the reference model, or its M and F averages are far-away. Likewise, a variable-wise decomposition of the unexplained part is also feasible. Other things being equal a variable's contribution is small if the respective coefficients in the different models are very close to each other.

## 4 RF and CART adapted to Blinder-Oaxaca

To generalize the Oaxaca-Blinder decomposition we need three models: one for the female, one for the male subsamples, and a reference for the pooled sample. OLS and Random Forest Regression are run on the male and female training samples, resulting in prediction functions $P^M$ and $P^F$, respectively. We also ran reference regressions (labelled by $P^R$) on the training sample.

These prediction functions are then applied to test samples, again divided into a male and a female subset to check the generalizability of our estimates. The following identity holds:

$$av(y_M) - av(y_F) = av\left(P^M(M)\right) - av\left(P^F(F)\right) + bias$$

where the arguments M and F refer to the identity of the subsamples, $av(y_M) - av(y_F)$ is the difference of average male and female log wages (the raw gender pay gap) and $av\left(P^M(M)\right) - av\left(P^F(F)\right)$ is the predicted average gender pay gap. We will study the following decomposition:

$$av\left(P^M(M)\right) - av\left(P^F(F)\right) = \left[av\left(P^R(M)\right) - av\left(P^R(F)\right)\right] +$$

$$\left[av\left(P^M(M)\right) - av\left(P^R(M)\right)\right] + \left[av\left(P^R(F)\right) - av\left(P^F(F)\right)\right],$$

where the first term on the right-hand side is the explained part, and the rest is the unexplained part. Notice that the unexplained part gives for each individual an unexplained residual, whereas in the case of the explained part only the averages can be compared. 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 explain fully the raw gap. If the "unexplained" gaps were non-zero, then we would think that the wage-setting mechanism conditional on our predictors works in an apparently discriminating manner.

To interpret the results, the individually estimated unexplained parts data can be modelled by CART, with the same explanatory variables. The CART output can be used to separate segments of the input space where the unexplained part is particularly large, or small.

## 5 An example

The dataset used for this example 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 of each year) on workers' age, gender, occupation, earnings (disaggregated into regular pay and irregular bonuses), types of contract, and whether the worker was hired recently. It also contains information about the employer (sector, region, settlement type, size of employment). The sampling procedure of 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 used the logarithm of the gross monthly wage, including regular wage and bonuses, for the response variable. The dataset was restricted to employees working full-time in the private sector. Calculations were carried out for the year 2008. The training sample contained 60 000 annual observations, and the rest made up the test sample. The raw gap was 0.1164 log-points.

Table 1 lists the predictors.

Table 1

| Name | Unit |
|---|---|
| Age | Years |
| Tenure | Months (at current employer) |
| Education* | 9 levels<br>(levels 8 and 9 are College and University) |
| New entrant dummy | 0: no, 1: yes |
| Share of foreign property* | 4 levels |
| Share of state property* | 4 levels |
| Firm size | Number of employees |
| Settlement | Capital city (Budapest)<br>Town<br>Other |
| Region | 7 categories |
| Sector | NACE Rev. 2 - 2 digits |
| Collective agreement on firm level | 0: no, 1: yes |

* This variable is refered to be ordered in Random Forests and in CART algorithm, too.

| | |
|---|---|
| Collective agreement on sector level | 0: no, 1: yes |
| Collective agreement with more employers but not on sector level | 0: no, 1: yes |
| Share of white collar employees in enterprise | percentage |

Source: Hungarian Wage Survey

Several reference models have been proposed in the literature: we opted for the Neumark, 1988 variety where gender is not included as an explanatory variable.

The Random Forest algorithm requires the setting of several parameters. In particular, our forests contained 500 trees each where OBB error seemed small enough (see in Appendix). To control for the growth of individual trees we set the minimum node-size parameter at 5 (default setting), 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).

## Results

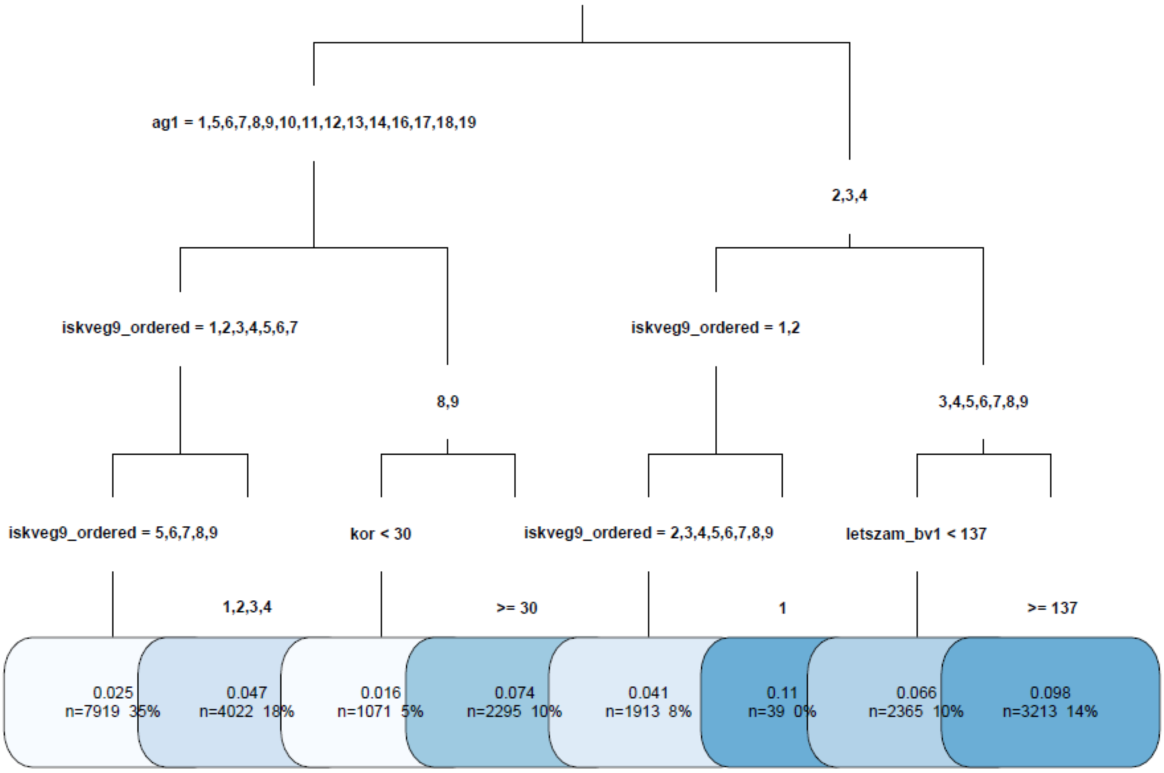
Table 2

| | Raw gap in log points | Explained (%) | Unexplained (%) | Bias (%) |
|---|---|---|---|---|
| Training data | 0.090 | -20 | 121 | -1 |
| Test data | 0.089 | -40 | 144 | 4 |

The first row of Table 2 shows the results on the training data, and the second row on the test data set. The first column contains the raw differences in log points. The next three columns show the explained and unexplained parts, moreover the bias expressed as percentages of the raw difference. For instance -40 (second row, second column) means that according to the test data average wage differences explained by observable characteristics should have had the opposite sign (women's wages must have been higher than men's), having a size of 40 % of the raw gap.

Figure 1 shows the tree of depth 4 (i.e. the 8 level tree after exactly three cuts).

Figure 1



Source: own calculation

We can look for interesting classes of people on Figure 1. For instance, we can find the group with the highest average unexplained residual (0.11). This group is characterized by the following properties: the lowest possible level of schooling (under 8 finished years in primary school), and working in the following branches: mining, manufacturing, electricity. This group, however, amounts to less than 1 % of the whole sample. The second highest unexpected residual can be observed for a group with 14 % of the total (0.098). This group is made up by women who work in the same three sectors, but in firms with more than 137 employees, and have medium or higher level of education. At the other extreme we look for the group with the lowest average unexplained residual (0.014), its features are: younger than 30 years of age, college or university education, and working in the sectors different from the three sectors mentioned above.   These results suggest that “discrimination” – in the limited sense of the term used in this paper - can be a sectoral and large firm phenomenon.

## 6 Summary

In this note we derived the Blinder-Oaxaca decomposition with the help of the nonparametric Random Forest regression. In that case we do not obtain a variable-by-variable decomposition, but the average explained and unexplained parts can be readily interpreted in the usual manner. One slight change from the traditional method is that the raw difference is not exactly equal to the sum of the explained and unexplained parts, but, with an estimation method that fits well, the deviation must be insignificant. Insight into the structure of the relationships can be gained by analysis with a CART regression of the unexplained residuals, which identifies those subgroups that are most or least discriminated.

## 7 References

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353-7360.

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

Breiman, L.–Friedman, J.–Stone, C. J.–Olshen, R. A. (1984): Classification and regression trees. CRC Press.

Grömping, Ulrike. 2009, "Variable importance assessment in regression: linear regression versus random forest." The American Statistician 63.4: 308-319.

Hastie, T., Tibshirani, R., Friedman, J., 2017. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer-Verlag,New York.

Howard, J., & Bowles, M. (2012). The two most important algorithms in predictive modeling today. In Strata Conference: Santa Clara.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. Electronic Journal of Statistics, 1, 519-537.

Jann, B., 2008. The Blinder–Oaxaca decomposition for linear regression models. The Stata Journal, 8(4), 453-479. http://dx.doi.org/10.1177/1536867X0800800401

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

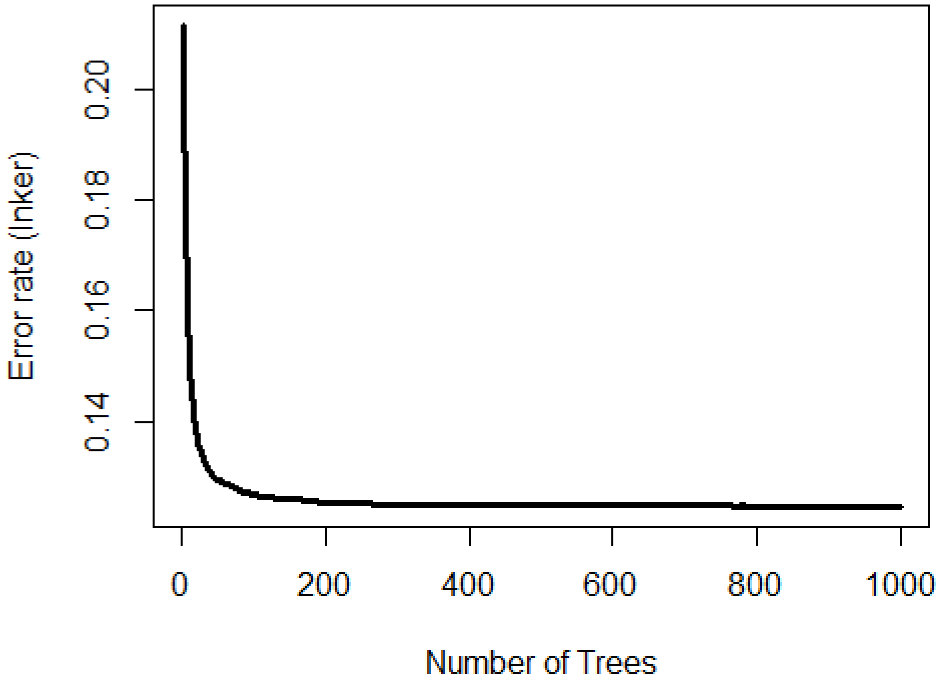
Varian, H. R., 2014. Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3-28. http://dx.doi.org/10.1257/jep.28.2.3

Wager, S., Athey, S., 2017. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228-1242. http://dx.doi.org/10.1080/01621459.2017.1319839

Appendix

Figure 2

OBB error in the case of Random Forest for female



Source: own calculation