INSTITUTE OF ECONOMICS
Centre for Economic and Regional Studies

Közgazdaság- és Regionális Tudományi Kutatóközpont
KÖZGAZDASÁG-TUDOMÁNYI INTÉZET

KRTK-KTI WORKING PAPERS | KRTK-KTI MŰHELYTANULMÁNYOK

# Heterogeneous wage structure effects: a partial European East-West comparison

OLGA TAKÁCS – JÁNOS VINCZE

**KRTK-KTI WP – 2023/05**

March 2023

https://kti.krtk.hu/wp-content/uploads/2023/03/KRTKKTIWP202305.pdf



INSTITUTE OF ECONOMICS, CENTRE FOR ECONOMIC AND REGIONAL STUDIES
EÖTVÖS LORÁND RESEARCH NETWORK (ELKH)
BUDAPEST

ABSTRACT
We estimate heterogeneous wage structure effects for country-pairs within the EU by the Causal Forest algorithm, then identify groups of workers with the highest and lowest discrepancies in terms of wage differentials. We find that, in the East-West comparison, age is the most consistently differentiating factor. People over 40 are most adversely treated in the East relative to the West, and especially those who have no tertiary education and work in small or medium-sized enterprises.

JEL codes: J31 F66 C14
Keywords: Wage structure effects, Generalized Random Forest Regression, Conditional average treatment effects, Wage convergence in the EU

Olga Takács
Corvinus University of Budapest
e-mail address:tackyolgi@gmail.com

János Vincze
Corvinus University of Budapest and
Centre for Economic and Regional Studies, Hungary
e-mail address: vincze.janos@krtk.hu

# Heterogén bérstruktúra hatások: egy Kelet-Nyugat összehasonlítás

TAKÁCS OLGA – VINCZE JÁNOS

ÖSSZEFOGLALÓ

Heterogén bérstruktúra hatásokat becslünk keleti-nyugati országpárokra Európán belül. Az Oksági Erdő algoritmus segítségével igyekszünk olyan munkapiaci szegmenseket azonosítani, amelyekre a Nyugat-Kelet százalékos bérkülönbségek robusztusan a legnagyobbak. Azt találjuk, hogy az életkor az az ismérv, ami a legkarakterisztikusabban megkülönbözteti a keleti béreket a nyugatiaktól, az idősebbek (40 év felettiek) relatív helyzete keleten különösen kedvezőtlen. Kevésbé konzisztensen, de az átlagnál nagyobb relatív hátrány jellemzi azokat a kevésbé képzett munkavállalókat is, akik kis vagy közepes vállalatban dolgoznak a non-tradable szektorban.

JEL: J31 F66 C14

Kulcsszavak: bérstruktúra hatások, Véletlen Erdők, feltételes átlagos kezelési hatások, bérkonvergencia az EU-ban

## I Introduction

Wage disparities are significant within the EU (Drahokoupil-Piasna, 2018). The present paper studies the heterogeneity of wage differentials (log wage differences) specifically in the East-West comparison. This comparison is important for various reasons. The lack of wage convergence implies incipient reallocation, either of capital or labour, and is considered as a problem to social cohesion. To interpret general wage differentials, we need to know more about their structure. For instance, if the wage differential is particularly high for older employees, the chance of re-allocating labour across countries is more meagre than if it were high mainly for younger workers.

Depending on the vantage point, our main question can be formulated in terms of relative (dis)advantages. A relative advantage from the vantage point of an employee is a disadvantage from that of the employer. As a manner of speech, we will take the employee's point of view; thus, we say, for instance, that in Hungary, older workers have substantial disadvantages with respect to younger workers compared with Belgium, meaning by it that the old/young wage differential in Belgium is much higher than in Hungary.

Heterogeneous wage discrepancies can be characterised in various ways. Our approach is to identify heterogeneous average wage structure (HWS) effects, i.e. we estimate expected wage differentials conditioned on multiple observable covariates. The traditional approach is to derive a variable-by-variable characterisation by estimating or charting the effect of individual covariates. The variable-wise decomposition of the wage structure effect in Oaxaca-Blinder decompositions is an instance of this method. However, the mutual dependence between covariates and the presence of non-linearities put limits to the meaningfulness of this approach. Therefore, rather than focussing on the effect of individual variables, we attempt to identify heterogeneity by pointing out those subsets of the covariate space where heterogeneity is salient. Indeed, this group-oriented description of heterogeneity is more natural for many practical purposes. When political or other decisions are formulated, one frequently considers subsets (groups) of individuals rather than individual properties. Policy eligibility criteria implicitly reflect heterogeneous treatment, which is more the rule than the exception in policy-making. Audience segmentation is omnipresent in marketing and health policies, for instance medical recommendations are defined for "segmented audiences", for instance, by age, gender, previous health conditions (Slater 1996.). Of course, segments are defined by a conjunction of ranges of relevant variables. The point is that for identifying the group that runs the highest health risk, one needs not quantify the risk born by individual variables separately.

We estimate expected wage differentials between country- pairs. Thus, our HWS effects are similar to CATE (conditional average treatment effects) in the treatment effects literature (Künzel et al., 2019; Athey-Imbens, 2016; Jacob, 2021). Though they may be thought of as "causal effects", we do not interpret them as such. Traditionally, CATE was estimated by non-parametric regression, but machine learning (ML) methods are appropriate tools for this purpose as well (Wager-Athey, 2018). To estimate the HWS effects, we applied random forest methodologies. Our preferred one, the causal forest algorithm of Athey et al. (2019), was specifically developed for such a purpose. Looking for heterogeneity means that our covariates are not mere controls whose only job is to assure that the parameter of interest (the average wage structure effect) be estimated correctly. As the HWS effects is a function of several dimensions their characterisation is not straightforward, they need to be processed to provide digestible information. A simple method is to apply a classification algorithm to the HWS effects by growing shallow surrogate trees (Molnar, 2020), and to select the most and least affected leaves (subgroups). A more sophisticated technique is to apply the policy-tree algorithm of Athey-Wager (2021), which divides the covariate space into disjoint groups via a decision tree, by optimising the division and separating the most and least affected groups by exhaustive search. Another way is to find the covariate subset with top and bottom estimated effects and calculate the average characteristics of these endogenously defined groups. This is the classification analysis proposed in Chernozhukov et al. (2018).

Our data come from the European Structure of Earnings Survey (SES), a large survey providing comparable microdata on earnings, employees', and their employer. (We use the 2018 wave in this paper.) Its most attractive feature is the application of a common questionnaire for all countries and the fact that enterprises provide the earnings data. It has been used in several studies that compared wage structures among European countries (Simon, 2010; De Caju et al.,2010)).

Our strategy is to consider pairwise contrasts between Eastern and Western countries. Naturally, we search for robust results by finding submarkets that are extremely affected by the East-West gradient in many comparisons, but we will also notice country specificities that may be interesting in their own right.

The rest of the paper is structured as follows. The next Section connects the paper with the literature. Section 3 describes the empirical methodologies, and Section 4 the data. Section 5 reports the results. Section 6 interprets our findings and points out questions they raise.

## II Related literature

### 1 East-West wages

A paper that addresses issues close to ours is Pereira-Galego (2018). It studies heterogenous wage differences among a couple of EU countries, but it focuses on the decomposition of unconditional quantiles. The authors used the SILC-EU data set, since their main target was to inform migration policy.

Drahokoupil-Piasna (2018) highlights differences along the East-West division; it quantifies both wage structure and composition effects. The methodology is multiple linear regressions, and heterogeneous effects are estimated in predefined subpopulations, a method ubiquitous in the older literature on heterogeneous effects. One of the main conclusions is that higher skills are relatively under-priced in Central and Eastern Europe.

SES data have been used in several international comparisons. Simon (2012) estimated common wage equations for nine countries but focussed on a single aspect of heterogeneity, the differences in the gender wage gap. Studies based on SES data with a similar goal are Boll-Lagemann (2019), and Leythienne-Ronkowski (2018). In a previous paper (Simón, 2005) Simon used a former edition of the SES to study international differences in employers' effects on wages. (See also Fernandez et al. (2004) for a similar approach but concentrating on low wage labour.) Simón (2010) also addressed wage inequality and its sources in international comparison, by searching for pairwise country discrepancies.

Magda et al. (2012) employed SES data to study the relationship between collective agreements, wages and restructuring in three transition countries by exploiting the idea that long tenure can be taken as an indicator that a firm had existed in the pre-transition period. Du Caju et al. (2010) explore the time dimension of the SES, and concentrate on the deviations between and explanation of industry wage differentials.

### 2 CATE estimation

As Fortin et al. (2011) pointed out, the wage structure effect can be regarded as a treatment effect, without the causal interpretation. Recently, a large literature developed in machine learning with the purpose of creating „meta-learners” whose output is the measurement of heterogeneous treatment effects (Künzel et al., (2019). Theoretical studies appeared also in the econometric literature (Athey-Wager, 2021) with the same purpose, and we make use of the methods developed in Athey et al. (2019). Knaus, et al. (2021) is a comparative study on the performance of several methodologies.

The Machine Learning literature (e.g. Molnar, 2020) is aware of the importance to interpret or process the usually non-parametric estimates. There are several suggestions how to do it, and we partly follow Molnar (2020) by fitting shallow regression trees on the HWS effects with the 'ctree' algorithm of Hothorn et al. (2006). A more sophisticated version of the idea to characterize the HWS effects by a decision tree is the policy-tree framework of Wager-Athey (2021). Also, we apply the CLAN methodology proposed in Chernozhukov et al. (2018), whose purpose is „to discover *ex post* whether there is any relevant heterogeneity in treatment effect by covariates". Papers methodologically similar to ours now galore (Deryugina et al., 2019; Davis-Heller, 2020; Murakami et al., 2020; Bertrand et al., 2022). Baiardi-Naghi (2021) revisits a number of studies using ML techniques to estimate CATE, and shows that interesting novelties come up from these exercises.

## III Methodologies

### 1 HWS effects estimates

We can regard our fundamental estimands (the HWS effects) as conditional average treatment effects (CATE), where we consider, arbitrarily, the Western country employees as the treated sub-population. Let Y denote log wages, and X a multidimensional set of primary covariates common to both East and West, and we assume common support. We estimate

$$\mathrm{HWS}(x) = E(y|x, T = 1)) - E(y|x, T = 0),$$

where $E(|)$ denotes conditional expectations, with T=1 for a Westerner and T=0 for an Easterner. The treatment effect interpretation relies on the assumption of un-confoundedness, which is questionable, but we strive to get robust conclusions that are probably not sensitive to it.

The covariate data available in the SES database are essentially ordinal, qualitative variables; thus, the set of possibly distinct explanatory variable combinations is finite. Therefore, the functional form of the conditional expectation function is known *a priori*, there is no difference between parametric and non-parametric specifications, and we can only choose the estimation methodology. We favour a tree-based technique, as these can naturally incorporate ordinal information. In particular, our first choice is the Causal Forest algorithm of Athey et al. (2019) (see also Athey-Wager (2019) for an application), which was explicitly developed as an adaptation of the random forest methodology fit for this task. Nonetheless, we made calculations with another random forest type algorithm to check the robustness of our findings.

As Wager-Athey (2018) emphasised, one of the possible applications of machine learning (ML) methods in economics is for problems where our main interest is in estimating a (possibly) nonlinear conditional expectation function (CEF). They assert that "anyplace in traditional econometrics where a kernel function might have been used, ML methods that perform better than kernels in practice may be substituted". Random Forest (RF) is a tree-based statistical learning algorithm that has been applied in many disciplines. Wager and Athey (2018) described RF regression as similar to other traditional non-parametric regression methods (e.g. k-nearest-neighbour algorithms), as it delivers some weighted average of "nearby" points as the prediction. However, it has the advantage that both the weights and the proximities are determined in a data-driven way.

RF was introduced in order to rectify the overfitting tendencies of single-tree based methods. In an RF regression, one grows many suboptimal regression trees, and the RF prediction is calculated as an average of the individual trees' predictions. Each tree is grown on a random subsample (or on a bootstrap sample), and at each node (possibly) only a random subset of explanatory variables is considered for a split. The main advantage of RF seems to be that the random and restricted manner of branch formation in individual trees achieves de-correlation among constituent trees.

The original RF algorithm of Breiman (2001) aimed at prediction accuracy, a goal that requires exploiting the bias-variance trade-off. Predictive performance (evaluated on independent data sets) may not imply other desirable statistical properties like consistency or the ability to conduct valid tests. Intensive research has led to the development of new types of RF algorithms that solve these problems, and yield consistent estimates and asymptotic normality, allowing to evaluate the precision of the estimates. The generalised random forest (grf) algorithm of Athey et al. (2018) implements changes with respect to Breiman’s RF with a view towards consistency. One modification embodied in it entails penalising unbalanced splits via two parameters to discourage the forming of very different sized child nodes. The most crucial difference is, probably, "honesty". Honesty separates splitting from estimation by dividing the (sub)sample into two parts. One of them is employed for placing the split, and the other one for estimation in the leaves. These changes contain the tendency of the original algorithm to lead to overfitting.

Either using Breiman's RF or the GRF there exist several possibilities to estimate the HWS(x) function. In the case of an S-learner one would estimate a regression on each of the combined East-West country-pair samples and then plug in the empirical counterparts for each observation in the sample. This is called an S learner of the CATE since it is based on a "single" sample. A problem with the S-learner is that the base learner aims to minimize the mean squared error of the prediction, and not the mean squared error of the treatment effect. Of course, the individual treatment effects are not observed, therefore the CATE estimation must somehow circumvent this problem. The Causal Forest of Athey et al. (2018) devises a way to achieve this. Here the splits are chosen to maximise the "causal" effect directly. In a tree-based algorithm, the general idea behind splitting is that the new nodes have to be most different in terms of the target, which in a prediction-oriented exercise is the predicted variable, while in an effect-oriented study, it is the heterogeneous causal effect. As the latter is not directly observable, the CF algorithm estimates the causal effect at each node, i.e. on a series of smaller and smaller subsamples by a simple orthogonalized regression. This algorithm is, also, an instance of the "Generalized Random Forests" (GRF) family.

## 2 Processing the HWS estimates

By definition different profiles (covariate vectors) are assigned different HWS effects. The resulting non-parametric function must be characterized in some interpretable way. An obvious starting point is to check whether there exists significant heterogeneity, i.e. the function is not a mere constant. We test for it in a way suggested in Chernozhukov et al. (2018), by projecting an unbiased signal of the HWS effects on the estimated effects. If the projection parameter is significant we can be confident that our estimate indeed captures important heterogeneity.

### 2.1 Surrogate trees

Regression-trees have been employed in many areas to carry out a group-oriented data analysis in a supervised learning context. Single "greedy" regression-trees are generally classified as well-interpretable ML algorithms with not very good predictive performance. This contrasts them to RF algorithms that are based on a randomized large set of (not so greedy) regression trees, and are considered excellent predictors without straightforward interpretability. The recent literature suggests that predictive models based on RF (or other powerful) predictors can be made interpretable if we fit shallow regression-trees on the model's predictions. This is sometimes called a global surrogate analysis, where one tries to approximate the results of a not easily interpretable model with a highly interpretable one, in our case a regression tree. For the possibilities among regression tree algorithms, we followed Molnar (2020), and utilized the 'ctree' algorithm of Hothorn et al. (2006). ‘Ctree’ improves on the CART algorithm, which suffers from overfitting and biased variable selection, the latter because variables with more values have a better chance of being selected for splitting. ‘Ctree’ solves both problems by using statistical hypothesis tests to determine splits. First, it tests the multiple hypotheses that all covariates are independent of the response at each node. If the null is accepted, the algorithm stops. If it is rejected, then single variable tests for independence of each covariate are computed, and the p-values compared. An attractive feature of ‘ctree’ is that tests are distribution free permutation tests. The variable with the lowest p-value (indicating the strongest relationship with the response variable) is chosen, avoiding thereby the selection bias. After picking the splitting covariate, ‘ctree’ uses the same method as CART to find the split point.

For the sake of interpretability, we grow shallow conditional inference trees, of depth 3. This choice results in – at most – eight distinct submarkets, of which we can select the ones with the largest and smallest average HWS effects. The splits leading to these extreme groups can give a first impression of what submarkets can be taken as good candidates for discriminative wage setting, or, alternatively, what the main drivers of heterogeneity are.

### 2.2 Finding the best “policy tree”

Though a shallow regression tree computed by ‘ctree’ avoids the variable selection bias, it is still a greedy algorithm that cannot look forward. This situation can be amended by searching for the best policy tree of Athey-Wager (2021), which assigns each x either to be treated or untreated, in the form of a decision tree. The ‘policytree’ algorithm conducts an exhaustive search in the space of decision trees of a certain depth to select the best one., thus it has foresight, too. To be best amounts to producing the highest possible average effect while restricting the assignment rule to decision trees. In Athey-Wager (2021) it is argued that this constraint is sensible for practical purposes, as eligibility is usually defined by a sequence of criteria in policy contexts. The ‘policytree' algorithm makes use of propensity scores to calculate average effects (doubly robust estimation). In a medical application it is not evident that treatment is beneficial for all thus the policy tree naturally contains subgroups for whom the no-treatment proposal is optimal. As our goal is not treatment assignment but the separation of groups with considerable relative advantages and disadvantages, we have to normalise the HWS effects by deducing the average HWS effect from each HWS (x) estimate.

### 2.3 Classification analysis (CLAN)

Chernozhukov et al. (2018) suggested a model agnostic analysis with the express purpose of making sense of machine learning methods by focussing on "grosser" features of the estimated heterogeneous effects. The CLAN (classification analysis) methodology consists in finding average characteristics of the most and least affected observational units as defined by the predictor of the conditional wage structure effects. We compute estimates of several functions of the covariates belonging to the highest and lowest 20 % in terms of HWS effects. All these statistics are proportions, i.e. estimates of probabilities.

Chernozhukov et al. (2018) offers a special estimation and inference methodology (VEIN) that works even with possibly inconsistent ML estimates, and produces asymptotically valid confidence intervals. This methodology is based on repeated data splitting and honest estimation.

1. Split the sample randomly into an A(uxiliary) and a M(ain) subsample.
2. Construct the CATE estimator on the A sample,
3. Predict the CATE on the M sample, then determine the extreme groups in the M sample, and compute the estimates of the proportions together with upper and lower bounds of confidence intervals,
4. Calculate the difference between the most and least affected groups, together with confidence intervals.
5. Repeat from 1 to N times to obtain N point and confidence interval estimates.
6. The final estimate is the median of the N point estimates and the medians of the lower and upper bounds of the confidence intervals.

If the elementary confidence intervals have a level 1-alpha then the implied confidence level of the final estimate is approximately 1-2alpha, as we have to take into account both estimation and sample splitting uncertainty.

## IV Data

The European Structure of Earnings Survey (SES) is an extensive survey providing comparable microdata on earnings, employees' individual characteristics, and their employer. We use the 2018 wave in this paper. Our target variable is (log) gross hourly earnings that refer to the wages and salaries earned per hour paid in the reference month (October 2014) before any tax and social security contributions are deducted. Wages and salaries include overtime pay, shift premiums, allowances, bonuses and commissions. SES data are collected in enterprises with at least ten employees operating in all areas of the economy except public administration and agriculture. The survey is carried out on a sample of plants selected by stratified random sampling, while within plants, a random sample of employees is chosen. Its most attractive feature is the application of a common methodology for all countries, and the fact that earnings data are provided by enterprises.

We dropped the Health and Education sectors discarded observations for part-time workers and apprentices, the 14-19 age group and those in the highest percentile of earnings. Table 1 contains the description of covariates.

Table 1 List of covariates

| Name | Unit |
|---|---|
| Age | Years (5 levels) |
| Tenure | Years of employment at current employer (8 levels) |
| Education | ISCED codes (4 levels) |
| Occupation | ISCO main groups (8 levels) |
| Firm size | Number of employees (3 levels) |
| Sector | Tradable – Non-tradable |
| Majority ownership | State – Private |
| Gender | Female – Male |

Notes: Age: 20-29 years, 30-39 years, 40-49 years, 50-59 years, 60 + years. Tenure: new entrant, 1-5 years, 6-10 years, 10-15 years, 16-20 years, 21-25 years, 26-30 years, 30 + years. Education: level 1 basic education, level 2 secondary education, level 3 tertiary education up to 4 years, level 4 tertiary education more than 4 years. Occupation: ISCO main groups, except 6 and 0. Firm size: less than 50 (small), 51-250 (medium), 250 + (large). Sector: tradeable, NACE 2 sectors B,C, D, non-tradable the rest except A, O, P, Q.

In this investigation Eastern countries include the Czech Republic, Hungary, Poland and Slovakia, while Belgium, Denmark, France, Italy, the Netherlands and Norway belong to the

West, At each country-pair estimation, we used reduced samples by randomly disposing of observations to obtain an initially balanced sample.

## V Results

We imposed the common support assumption for every country pair by estimating propensity scores by ‘GRF’, and excluding the observations with an estimated score below 0.1 or above 0.9, as suggested in Austin (2009). The CF algorithm includes a heterogeneity check based on Chernozhukov et al. (2018), which shows that the WS effects are significantly heterogeneous in each comparison and the CF algorithm captures well the heterogeneity.

### 1. Conditional inference trees

The target variable is the estimated HWS effects, and the covariates are the same as in the CF algorithms. The significance level was set at 0.001, and the joint test used the Bonferroni adjustment. The results are summarised in Table 1. The least (most) affected leaves are those that have the smallest (largest) average score in the respective tree. In each cell in Table 1 we indicate the corresponding splits. When only one variable appears in a cell, both splits were done with respect to that variable.

From Table 1.a one can see that the most frequent joint features for the smallest effects usually involve age (22 of 24 cases). The usual partners of age include occupation, education and firm size. In Table 1.b age dominates again (19 appearances) and the accompaniment is roughly the same.

Table 1.a Leaves with smallest effects

| | Czech Republic | Hungary | Poland | Slovakia |
|---|---|---|---|---|
| Belgium | age 40-<br>male | LE<br>age 40- | age 40-<br>male | age 40-<br>male |
| Denmark | ISCO 12<br>age 40- | ISCED 34<br>LE | age 40-<br>ISCED 34 | ISCO 15<br>age 40- |
| France | age 50-<br>LE | age 40-<br>LE | LE<br>age 40- | LE<br>age 50- |
| Italy | ISCO 13<br>age 40- | age 40-<br>ISCO 14 | age 40-<br>ISCO 12 | age 40-<br>private |
| Netherlands | ISCO 12<br>age 40- | age 40-<br>LE | age 40-<br>MLE | age 30- |
| Norway | ISCO 13<br>age 40- | ISCED 34<br>age 40- | ISCED 34<br>LE | ISCO:1-3<br>age 40- |

Table 1.b.: Leaves with largest effects

| | Czech Republic | Hungary | Poland | Slovakia |
|---|---|---|---|---|
| Belgium | age 50+ | age 40+<br>ISCO 4-9 | age 40+<br>female | age 50+ |
| Denmark | ISCO 4-9<br>SME | SME<br>ISCED 12 | ISCED 12<br>SE | age 40+<br>LE |
| France | age 60+ | age 40+<br>SME | age 40+<br>SME | age 50+<br>ISCO 12 |
| Italy | age 50+ | age 50+ | age 50+ | age 50+<br>public |
| Netherlands | age 50+ | age 40+<br>SME | age 40+<br>SE | age 40+<br>ISCED 4 |
| Norway | ISCO 5-8<br>age 40+ | SME<br>ISCED 12 | ISCO 3-8<br>ISCED 12 | age 50+ |

Note:  SE stands for small, SME for small or medium-size, LE for large, MLE for medium or large enterprise.

These tables strongly suggest that the submarkets with the highest wage structure effects (the most relatively disadvantaged groups) include older people who are less educated, have simpler jobs and are employed by smaller enterprises. In contrast, the submarkets with the largest relative advantage contain highly qualified, younger people with professional or managerial jobs in larger enterprises.

## 2. Optimal policy trees

Table 2 shows the leaves of the optimal policy trees of depth 2. Cells in Table 2 must be understood as submarkets for which the treatment is the “recommendation". In other words these are the most relatively disadvantaged groups in the Eastern countries.

Table 2: Policy trees: groups going West

| | Czech Republic | Hungary | Poland | Slovakia |
|---|---|---|---|---|
| Belgium | ISCO 13, age 50+<br>ISCO 49 age 40+ | SME,age 40+<br>LE, age 50+ | F, age 30+<br>M, age 50+ | Pub, age 30+<br>Pri, age 40+ |
| Denmark | ISCO 13 ,age 50+<br>ISCO 49 age 30+ | SME, ISCED12<br>LE, ISCO 69 | Pub, SM<br>Pri ISCED 12 | ISCO15, age 40+<br>ISCO69, ISCED13 |
| France | SME<br>LE, age 50+ | SME, tenure 0+<br>LE, age 50+ | SME, age 30+<br>L, age 50+ | NT, age 40+<br>T, age 50 + |
| Italy | F, age 40+<br>M, age 50+ | ISCED12, age 50+<br>ISCED34, Pub | SME, age 40+<br>LE, age 50+ | NT, age 40+<br>T, age 50 + |
| Netherlands | Pub, NT<br>Pri, age 40+ | SME, age 40+<br>L, age 50+ | Pub, SME<br>Pri, age 40+ | Pub, NT<br>Pri, age 40+ |
| Norway | ISCO 13, age 50+<br>ISCO 49, ISCED 12 | SME, ISCED12<br>LE, ISCO 79 | age 30-, ISCO 79<br>age 30+, ISCO 16 | ISCO 15, age 40+<br>ISCO 69 |

Note: Pub stands for publicly owned, and Pri for privately owned enterprise. NT: non-tradable, T: tradable.

Age appears in 21 cells, and it is invariably suggested that older people “Go West”.

## 3. The CLAN

For the sake of the CLAN exercise, we dichotomised all covariates and presented estimates for one category per covariate to ease interpretation. The list of baseline categories which are coded as taking value 1 are as follows: Gender=Female, Education basic or secondary, ISCO 1-3 (Managers and Professionals), Age less than 40 yrs, Tenure more than 20 yrs, Ownership=State, Size less than 250, Sector=Tradable. For each Eastern country (see Tables 3.1-3.4) we report differences in proportions (most affected percentage minus least affected percentage), where least (most) affected stands for the lower (upper) 20 percentiles of the estimated wage structure effects. A positive coefficient in the columns implies that the respective baseline category has a "relative disadvantage" in the corresponding Eastern country. Yellow cells indicate that the parameter is not significantly different from 0 at the 10 % level. Starred red entries draw attention to those numbers that are above 0.33.

### 3.1 Hungary

| | Female | ISCED12 | ISCO13 | age 40- | tenure 10+ | Public | SME | Tradable |
|---|---|---|---|---|---|---|---|---|
| Belgium | 0.149 | 0.489 * | -0.475 * | -0.835 * | -0.054 | *0.012* | 0.491 * | -0.062 |
| Netherlands | -0.171 | 0.339 * | -0.282 | -0.876 * | -0.279 | 0.037 | 0.699 * | -0.132 |
| Italy | -0.046 | 0.474 * | -0.188 | -0.766 * | -0.390 * | 0.375 * | 0.356 * | -0.174 |
| France | -0.011 | 0.446 * | -0.110 | -0.554 * | -0.251 | 0.139 | 0.656 * | -0.256 |
| Denmar | -0.242 | 0.818 * | -0.717 * | -0.459 * | -0.041 | -0.049 | 0.674 * | -0.057 |
| Norway | -0.213 | 0.725 * | -0.522 * | -0.458 * | -0.003 | 0.048 | 0.520 * | -0.091 |

### 3.2 Czech Republic

| | Female | ISCED12 | ISCO1 | age 40- | tenure 10+ | Public | SME | Tradable |
|---|---|---|---|---|---|---|---|---|
| Belgium | 0.549 * | 0.212 | -0.354 * | -0.885 * | -0.099 | 0.067 | 0.356 * | -0.311 |
| Netherlands | 0.167 | -0.065 | -0.279 | -0.847 * | -0.473 * | 0.044 | 0.257 | -0.011 |
| Italy | 0.329 | 0.267 | -0.191 | -0.886 * | -0.506 * | 0.303 | 0.120 | -0.188 |
| France | 0.313 | 0.362 * | -0.328 | -0.696 * | -0.412 * | 0.212 | 0.381 * | -0.016 |
| Denmark | 0.166 | 0.621 * | -0.814 * | -0.604 * | -0.161 | 0.039 | 0.433 * | 0.155 |
| Norway | 0.025 | 0.570 * | -0.814 * | -0.541 * | -0.089 | 0.107 | 0.286 | 0.276 |

### 3.3 Slovakia

| | Femal | ISCED12 | ISCO13 | age40- | tenure10+ | Public | SME | Tradable |
|---|---|---|---|---|---|---|---|---|
| Belgium | 0.481 * | -0.207 | 0.109 | -0.887 * | -0.183 | 0.060 | 0.173 | -0.111 |
| Netherlands | 0.127 | -0.327 | 0.105 | -0.927 * | -0.392 * | 0.117 | 0.145 | -0.143 |
| Italy | 0.255 | 0.056 | 0.081 | -0.843 * | -0.437 * | 0.432 * | 0.104 | -0.294 |
| France | 0.253 | -0.186 | 0.283 | -0.676 * | -0.257 | 0.271 | 0.324 | -0.405 * |
| Denmark | 0.055 | 0.494 * | -0.523 * | -0.741 * | -0.159 | 0.113 | 0.391 * | 0.023 |
| Norway | 0.112 | 0.325 | -0.331 * | -0.626 * | -0.072 | 0.259 | 0.249 | -0.074 |

### 3.4 Poland

| | Female | ISCED12 | ISCO13 | age 40- | tenure 10+ | Public | SME | Tradable |
|---|---|---|---|---|---|---|---|---|
| Belgium | 0.545 * | 0.342 * | -0.336 * | -0.892 * | -0.206 | -0.01 | 0.411 * | -0.028 |
| Netherlands | 0.094 | 0.218 | -0.200 | -0.873 * | -0.436 * | -0.08 | 0.491 * | -0.050 |
| Italy | 0.226 | 0.368 * | -0.158 | -0.830 * | -0.532 * | 0.190 | 0.263 | -0.074 |
| France | 0.177 | 0.426 * | -0.222 | -0.679 * | -0.400 * | -0.00 | 0.621 * | 0.057 |
| Denmark | 0.005 | 0.839 * | -0.821 * | -0.519 * | -0.154 | -0.14 | 0.581 * | 0.119 |
| Norway | -0.018 | 0.812 * | -0.802 * | -0.402 * | -0.090 | -0.06 | 0.343 * | 0.266 |

Note: The causal forest estimates underlying this table were calculated with the R package 'grf', using default settings. The VEIN method was applied with 100 random splits.  Least (most) affected: lower (upper) 20 percentiles by estimated mean wage structure effects.

Tables 4.1-4.8 summarise the most important qualitative findings. In these tables ++ (--) indicate that the change in proportions, i.e. proportion of most affected minus proportion of least affected, is more (less) than 33 (-33) percent. Otherwise, the + or - sign shows the sign of the difference, and 0 indicates that the difference is not significantly different from 0 at the 10 % level of significance.

Table 4.1 CLAN: gender

| | Czech Republic | Hungary | Poland | Slovakia |
|---|---|---|---|---|
| Belgium | ++ | + | + | ++ |
| Denmark | + | - | + | 0 |
| France | + | - | + | + |
| Italy | + | - | + | - |
| Netherlands | + | - | + | 0 |
| Norway | + | - | + | + |

The plus sign here means that there are more women in the most affected group. Gender does not appear as an important distinction between East and West, but there are hints that female labour has a robust relative disadvantage in Poland and the Czech Republic.

Table 4.2 CLAN: Education

| | Czech Republic | Hungary | Poland | Slovakia |
|---|---|---|---|---|
| Belgium | + | ++ | ++ | + |
| Denmark | ++ | ++ | ++ | ++ |
| France | ++ | ++ | ++ | + |
| Italy | ++ | ++ | ++ | + |
| Netherlands | + | ++ | ++ | + |
| Norway | ++ | ++ | ++ | ++ |

Educational achievement seems an important East-West watershed, though it is less clear-cut in the Czech Republic and Slovakia. Invariably less skilled people are relatively disadvantaged in the East.

Table 4.3 CLAN: Occupation

| | Czech Republic | Hungary | Poland | Slovakia |
|---|---|---|---|---|
| Belgium | - | -- | -- | - |
| Denmark | -- | -- | -- | -- |
| France | -- | - | -- | - |
| Italy | -- | - | - | - |
| Netherlands | -- | - | -- | - |
| Norway | -- | -- | -- | -- |

Occupation and education are obviously correlated, as jobs that carry higher decision-making responsibilities tend to be fulfilled by people of higher educational achievement. Jobs that needed higher skills are apparently relatively less disadvantaged in the East.

Table 4.4 CLAN: Age

| | Czech Republic | Hungary | Poland | Slovakia |
|---|---|---|---|---|
| Belgium | -- | -- | -- | -- |
| Denmark | -- | - | -- | -- |
| France | -- | -- | - | -- |
| Italy | -- | -- | -- | -- |
| Netherlands | -- | -- | -- | -- |
| Norway | - | - | - | -- |

More advanced age definitely implies relatively lower wages in the East. The comparison with Norway is the only one where this is not striking.

Table 4.5 CLAN: Tenure

| | Czech Republic | Hungary | Poland | Slovakia |
|---|---|---|---|---|
| Belgium | + | + | + | + |
| Denmark | + | - | 0 | + |
| France | + | + | + | + |
| Italy | + | + | + | + |
| Netherlands | + | + | + | + |
| Norway | - | - | - | - |

As tenure and age are obviously not independent, the result of age predicts that of tenure. A closer inspection of the CLAN tables shows that longer tenure softens the Eastern disadvantage of old age.

Table 4.6 CLAN: Control

| | Czech Republic | Hungary | Poland | Slovakia |
|---|---|---|---|---|
| Belgium | + | + | 0 | + |
| Denmark | - | + | - | + |
| France | + | + | - | + |
| Italy | ++ | ++ | + | ++ |
| Netherlands | + | 0 | - | + |
| Norway | 0 | + | - | + |

Ownership is not a particularly East-West dividing line. Italy and Poland seem to offer relative advantages to those employed in state-owned enterprises.

Table 4.7 CLAN: Size

| | Czech Republic | Hungary | Poland | Slovakia |
|---|---|---|---|---|
| Belgium | ++ | ++ | ++ | ++ |
| Denmark | ++ | ++ | ++ | + |
| France | + | ++ | ++ | + |
| Italy | + | + | ++ | + |
| Netherlands | + | ++ | ++ | + |
| Norway | + | ++ | ++ | + |

Size makes another clear-cut division; smaller enterprises seem to pay relatively less in the East.

Table 4.8 CLAN: Sector

| | Czech Republic | Hungary | Poland | Slovakia |
|---|---|---|---|---|
| Belgium | -- | - | - | -- |
| Denmark | + | - | + | + |
| France | - | - | - | - |
| Italy | - | - | - | - |
| Netherlands | + | - | - | - |
| Norway | + | - | + | - |

There seems a tendency to pay relatively more in the tradable sector in the East, which might be a consequence of the fact that firms are larger in this sector.

These tables confirm the central role of age: above 40 the wage disadvantage tends to be large. Tenure does not modify this observation, longer tenure is associated with larger disadvantages, but, of course, longer tenure correlates with older age. The role of enterprise size is also consistent in general; working in smaller companies in the East is tantamount to having a relative wage disadvantage. Higher education is robustly associated with less disadvantage, though the finer distinctions in educational achievement may not be fully comparable. Job responsibility is, of course, correlated with educational achievement and the numbers reflect this. There is an interesting phenomenon, detectable in detailed Tables (not reported here), that the Second main occupational group (professionals without particular managerial responsibilities) contains the less disadvantaged people in the East. Finally, working in the tradable sector brings fewer disadvantages.

The causal forest algorithm calculated our benchmark HWS effects estimates, and we checked whether the two S-learners (with Breiman's random forest and with GRF) would hold up essentially the same qualitative results. The two alternative methodologies reproduced the most important qualitative findings, though the size of the effects was, in general,smaller. In addition, we checked whether the imposition of the common support assumption makes any difference. Finally, we wanted to make sure that intra-country regional wage differences cannot result in heterogeneity accidentally. Therefore, we constructed new wages for each country by partialling out regional differences, i.e., equalising average regional wages country-wise. Results of both exercises did not show noticeable changes.

## VI Interpretation

Our results appear to be at variance with former findings in the literature. For instance, Pereira-Galego (2018) concluded that high positive wage structure effects provide incentives for Eastern European workers to migrate to the West (which is hardly questionable), and that the strength of the incentives are independent of the observed characteristics. Our paper bears on the issue of heterogeneity, too, but we approached it differently and obtained apparently contrasting results. The method of Pereira and Galego was based on influence function regressions and aimed at estimating differences at certain quantiles. It is possible that quantiles mix up subgroups (sub-markets), and makes heterogeneities fade away. Drahokoupil-Piasna (2018) finds that skills are comparatively lower in Eastern countries. As skills may improve with age our findings are not contradictory to this conclusion. However, we do not have a synthetic skill variable, rather we study its constituting colours distinctly.

Our findings may reveal a sorting effect. Possibly the relative (West-East) human capital differences are larger for those whose schooling and labour market experience fall mostly on the pre-1990 period, and older and less educated people are sorted with smaller and less efficient firms. Another possibility is that higher wage differentials for older and less educated people reflect their relatively weaker bargaining position in the East. Whichever explanation is correct, they indicate that more substantial potential incentives exist for those who cannot easily exploit them, a conclusion with the colour of equilibrium reasoning. However, for the sake of social cohesion, the fact itself is not laudable.

An interesting follow-up of the present investigation would be to study sectoral relative wage advantages. The Balassa-Samuelson effect is used to explain many closely related observed facts, like the positive correlation between income and the price level and the tendency of real exchange rate appreciation for developing countries. The basic argument rests on two assumptions: wages are higher in wealthier countries in the tradable sector, while the tradable-non-tradable relative wage is the same in all countries. Within country wage equalisation is not essential, the validity of the theory rests only on the assumption that the relative tradable-non-tradable wage ratio is the same in all countries. Our results may suggest that this is not the case within Europe.

**VII References**

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

Athey, Susan, and Stefan Wager. "Estimating treatment effects with causal forests: An application." *Observational Studies* 5.2 (2019): 37-51.

Athey, Susan, Stefan Wager. "Policy learning with observational data." *Econometrica* 89.1 (2021): 133-161.

Athey, Susan, Julie Tibshirani, and Stefan Wager. "Generalised random forests." *The Annals of Statistics* 47.2 (2019): 1148-1178.

Austin. P.C. Using the standardised difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics - Simulation and Computation*, 38(6):1228-1234, (2009). URL https://doi.org/10.1080/03610910902859574.

Baiardi, A.,Naghi, A.A., The value added of machine learning to causal inference: Evidence from revisited studies. *arXiv preprint arXiv:2101.00878* (2021).

Bertrand, Hussam, R., Rigol, N. and Roth, B.N., Targeting high ability entrepreneurs using community information: Mechanism design in the field. *American Economic Review*, *112*(3), pp.861-98. (2022)

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

Chernozhukov V, Demirer M, Duflo E, Fernández-Val I. Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomised Experiments, with an Application to Immunisation in India. *National Bureau of Economic Research*, No. w24678. (2018).

Davis, J.M. and Heller, S.B., Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. *Review of Economics and Statistics*, *102*(4), pp.664-677. (2020)

Deryugina, T., Heutel, G., Miller, N.H., Molitor, D. and Reif, J., The mortality and medical costs of air pollution: Evidence from changes in wind direction. *American Economic Review*, *109*(12), (2019), pp.4178-4219. ,

Du Caju, Philip, et al. Inter-industry wage differentials in EU countries: what do cross-country time varying data add to the picture? *Journal of the European Economic Association*, 8.2-3 (2010): 478-486.

Drahokoupil, Jan, and Agnieszka Piasna. What is behind low wages in central and eastern Europe? *Post-Communist Economies* 30.4 (2018): 421-439.

Fernández, M., Meixide, A., Nolan, B., & Simon, H. Low wage employment in Europe. *London, Pay Inequalities and Economic Performance Working Paper*, (2004

Fernández, Melchor, et al. "Low wage employment in Europe." *London, Pay Inequalities and Economic Performance Working Paper* (2004).

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. "Decomposition methods in economics." *Handbook of Labor Economics*. Vol. 4. Elsevier, (2011). 1-102.

Henrion, Thibaut, Javier Alcántara Ortega, and Denis Leythienne. Wages determinants in the European Union: Evidence from structure of earnings survey (SES 2014) data, *Eurostat*. (2020).

Hothorn, Torsten, Kurt Hornik & Achim Zeileis: Unbiased Recursive Partitioning: A Conditional Inference Framework, *Journal of Computational and Graphical Statistics*, 15:3, 651-674, DOI: 10.1198/106186006X133933, (2006).

Jacob, Daniel. CATE meets ML-The Conditional Average Treatment Effect and Machine Learning, *arXiv preprint arXiv:2104.09935* (2021).

Knaus, M.C., Lechner, M. and Strittmatter, A., Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, *24*(1), pp.134-161 (2021).

Knaus, M.C., Lechner, M. and Strittmatter, A., Heterogeneous Employment Effects of Job Search Programs A Machine Learning Approach. *Journal of Human Resources*, *57*(2), pp.597-636. (2022).

Künzel, Sören R., et al. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116.10 (2019): 4156-4165.

Leythienne, Denis, and Piotr Ronkowski. A decomposition of the unadjusted gender pay gap using Structure of Earnings Survey data. *Publications Office of the European Union*, (2018).

Magda, I., Rycx, F., Tojerow, I., & Valsamis, D.: Wage differentials across sectors in Europe: An East–west comparison 1. *Economics of Transition*, *19*(4), 749-769. (2011).

Magda, Iga, David Marsden, and Simone Moriconi. Collective agreements, wages, and firms' cohorts: Evidence from Central Europe. *ILR Review* 65.3 (2012): 607-629.

Molnar, Christoph. *Interpretable Machine Learning*. Lulu. com, (2020).

Mullainathan S, Spiess J. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* (2017); 31(2): 87-106.

Murakami, K., Shimada, H., Ushifusa, Y. and Ida, T., Heterogeneous Treatment Effects of Nudge and Rebate: Causal Machine Learning in a Field Experiment on Electricity Conservation. *International Economic Review*. (2020)

Pereira, João, and Aurora Galego. Inter-country wage differences in the European Union. *International Labour Review* 157.1 (2018): 101-128.

Simón, Hipólito. International differences in wage inequality: a new glance with European matched employer–employee data. *British Journal of Industrial Relations* 48.2 (2010): 310-346.

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

Simón, Hipólito. Employer wage differentials from an international perspective. *Economics Letters* 88.2 (2005): 284-288.

Slater, Michael D. "Theory and method in health audience segmentation." *Journal of Health Communication* 1.3 (1996): 267-284.

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