The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees and Forests*

Paolo Brunori, Paul Hufe & Daniel Mahler

January 14, 2023

Abstract

We propose the use of machine learning methods to estimate inequality of opportunity and illustrate that regression trees and forests represent a substantial improvement over existing approaches: they reduce the risk of ad-hoc model selection and trade off upward and downward bias in inequality of opportunity estimates. The advantages of regression trees and forests are illustrated by an empirical application for a cross-section of 31 European countries. We show that arbitrary model selection may lead to significant biases in inequality of opportunity estimates relative to our preferred method. These biases are reflected in both point estimates and country rankings.

JEL-Codes: D31; D63; C38
Keywords: Equality of Opportunity; Machine Learning; Random Forests

*__Brunori__ (corresponding author): International Inequalities Institute, London School of Economics. Postal Address: London School of Economics, Houghton Street, London, WC2A 2AE, UKU (paolo.brunori@unifi.it); **Hufe**: University of Bristol, IZA, CESifo (paul.hufe@bristol.ac.uk); **Mahler**: World Bank (dmahler@worldbank.org). We thank Chiara Binelli, Marc Fleurbaey, Torsten Hothorn, Niels Johannesen, Andreas Peichl, Giuseppe Pignataro, Dominik Sachs, Jan Stuhler, Dirk Van de gaer, and Achim Zeileis for useful comments and suggestions. Furthermore, we are grateful for the comments received from seminar audiences at Princeton University, the University of Perugia, the University of Essex, the World Bank, ifo Munich, the University of Copenhagen, Canazei Winter School 2018, the European Commission JRC at Ispra, the EBE Meeting 2018, IIPF 2018, and the Equal Chances Conference in Bari. Any errors remain our own.

1 INTRODUCTION

Equality of opportunity is an important ideal of distributive justice. It has widespread support in the general public and its realization has been identified as an important goal of public policy intervention (Cappelen et al., 2007; Alesina et al., 2018; Corak, 2013; Chetty et al., 2016). In spite of its popularity, providing empirical estimates of equality of opportunity is notoriously difficult. Next to normative dissent about the precise factors that should be viewed as contributing to unequal opportunities, current estimation approaches are encumbered by ad-hoc model selection that lead researchers to over- or underestimate inequality of opportunity.

In this paper we propose the use of machine learning methods to overcome the issue of ad-hoc model selection. Machine learning methods allow for flexible models of how unequal opportunities come about while imposing statistical discipline through criteria of out-of-sample replicability. These features serve to establish inequality of opportunity estimates that are less prone to upward or downward bias.

The empirical literature on the measurement of unequal opportunities has been flourishing since John Roemer's (1998) seminal contribution, *Equality of Opportunity*. At the heart of Roemer's formulation is the idea that individual outcomes are determined by two sorts of factors: those factors over which individuals have control, which he calls *effort*, and those factors for which individuals cannot be held responsible, which he calls *circumstances*. While outcome differences due to effort exertion are morally permissible, differences due to circumstances are inequitable and call for compensation.[1] Grounded on this distinction, inequality of opportunity measures quantify the extent to which individual outcomes are predicted by circumstance characteristics. They are usually calculated in a two-step procedure. First, researchers predict an outcome of interest from observable circumstances. Second, they calculate inequality in the distribution of predicted outcomes: the more predicted outcomes diverge, the more are circumstances associated with outcomes, and the more inequality of opportunity there is.

[1] The distinction between circumstances and efforts underpins many prominent literature branches in economics such as the ones on intergenerational mobility (Chetty et al., 2014a,b), the gender pay gap (Blau and Kahn, 2017) and racial differences (Kreisman and Rangel, 2015). For different notions of equality of opportunity, see Arneson (2018).

Current approaches to estimate inequality of opportunity suffer from biases that are the consequence of critical choices in model selection. First, researchers have to decide which circumstance variables to consider for estimation.[2] The challenge of this task grows with increasing availability of high-quality datasets that provide very detailed information with respect to individual circumstances (Björklund et al., 2012; Hufe et al., 2017). On the one hand, discarding relevant circumstances from the estimation model limits the explanatory scope of circumstances and leads to downward biased estimates of inequality of opportunity (Ferreira and Gignoux, 2011). On the other hand, including too many circumstances overfits the data and leads to upward biased estimates of inequality of opportunity (Brunori et al., 2019). Second, researchers must choose a functional form according to which circumstances co-produce the outcome of interest. For example, it is a well-established finding that the influence of socio-economic disadvantages during childhood on life outcomes varies by biological sex (Chetty et al., 2016; Dahl and Lochner, 2012). In contrast to such evidence, many empirical applications presume that the effect of circumstances on individual outcomes is log-linear and additive while abstracting from possible interaction effects (Bourguignon et al., 2007; Ferreira and Gignoux, 2011). On the one hand, restrictive functional form assumptions limit the ability of circumstances to explain variation in the outcome of interest and thus force a downward bias on inequality of opportunity estimates. On the other hand, limitations in the available degrees of freedom may prove a statistically meaningful estimation of complex models with many parameters infeasible.

This discussion highlights the non-trivial challenge of selecting the appropriate model for estimating inequality of opportunity. Researchers must balance different sources of bias while avoiding ad-hoc solutions. While this task is daunting for the individual researcher, it is a standard application for machine learning algorithms that are designed to make out-of-sample predictions of a dependent variable based on a number of observable predictors. In this paper, we use conditional inference regression trees and forests to estimate inequality of opportunity (Hothorn et al., 2006). Introduced by Morgan and Sonquist (1963) and later popularized by Breiman et al. (1984) and Breiman (2001), they belong to a set of machine learning methods that is increasingly integrated into the statistical toolkit of economists (Varian, 2014; Mullainathan

[2] Roemer does not provide a fixed list of circumstance variables. Instead he suggests that the set of circumstances should evolve from a political process (Roemer and Trannoy, 2015). In empirical implementations typical circumstances include biological sex, socioeconomic background, and race.

and Spiess, 2017; Athey, 2018). Trees and forests obtain predictions by drawing on a clear-cut algorithm that imposes only minimal assumptions about which and how circumstances interact in shaping individual opportunities. Thereby, they restrict judgment calls of the researcher and inform model specification by data analysis. As a consequence, they cushion downward bias by flexibly accommodating different ways of how circumstance characteristics shape the distribution of outcomes. Moreover, the conditional inference algorithm branches trees (and constructs forests) by a sequence of hypothesis tests that prevents the inclusion of noisy circumstance parameters. This feature reduces the potential for upward biased estimates of inequality of opportunity through model overfitting. Hence, regression trees and forests address the detrimental consequences of ad-hoc model selection in a way that is sensitive to both upward and downward bias in inequality of opportunity estimates.

To showcase the advantages of regression trees and forests we compare them to existing estimation approaches in a cross-sectional dataset of 31 European countries. We demonstrate that current estimation approaches overfit (underfit) the data which in turn leads to upward (downward) biased estimates of inequality of opportunity. These biases are sizable. For example, some standard methods overestimate inequality of opportunity in the Nordic countries while they underestimate the extent of inequality of opportunity in Germany and France. As a consequence, these countries appear close in terms of their opportunity characteristics. Hence, standard estimation approaches may yield misleading information about the level of inequality of opportunity in different societies to policymakers and the general public alike.

While we demonstrate the advantages of regression trees and forests for estimations of inequality of opportunity, they are not a panacea to empirical challenges in this literature. First, regression trees and forests cannot address one of the most relevant sources of downward bias in inequality of opportunity estimates: missing data on relevant circumstances. Second, although regression trees and forests are less prone to upward and downward bias, the remaining bias may nevertheless be substantial when samples are small. Therefore, we encourage applied researchers to exercise caution when estimating inequality of opportunity on data with small number of observations.

The remainder of this paper is organized as follows: section 2 gives a brief introduction to current empirical approaches in the literature on inequality of opportunity. Section 3 introduces conditional inference regression trees and forests, and illustrates how to use them in the context of inequality of opportunity estimations. An empirical illustration based on simulated data and the EU Survey of Income and Living Conditions is contained in section 4. In this section we also highlight the particular advantages of tree- and forest-based estimation methods by comparing them to the prevalent estimation approaches in the literature. Section 5 concludes the paper.[3]

2 EMPIRICAL APPROACHES TO EQUALITY OF OPPORTUNITY

Theoretical Set-up and Notation. Consider a population $\mathcal{N} = \{1, ..., N\}$ and an associated vector of non-negative incomes $y = (y_1, ..., y_N)$. Income is determined by two sets of factors: *circumstances* beyond individual control and individual *efforts*. We define the $(P \times 1)$-vector $\omega_i \in \Omega$ as a comprehensive description of the circumstances of $i \in \mathcal{N}$. Analogously we define the $(Q \times 1)$-vector $\theta_i \in \Theta$ as a comprehensive description of the efforts that are exerted by $i \in \mathcal{N}$. The income generating function can be defined as follows:

$$y = d(\omega, \theta). \tag{1}$$

Based on the realizations of individual circumstances, the population can be partitioned into *types*. We define the type partition $\mathcal{T} = \{t_1, ..., t_M\}$, such that individuals are member of one type if they share the same circumstances: $i, j \in t_m \iff \omega_i = \omega_j$.

Measurement. Opportunity egalitarians are averse to inequalities that are rooted in circumstances, however, they are indifferent to inequalities that originate from individual effort exertion. In spite of the intuitive appeal of this idea, the literature has suggested a variety of formulations that differ in their precise normative content—see Ramos and Van de gaer (2016) for an overview. In this work we exclusively focus on *ex-ante utilitarian* measures of inequal-

[3] In a parallel paper, Blundell and Risa (2019) apply machine learning methods to the estimation of intergenerational mobility. In particular, they assess the completeness of rank-rank estimates of intergenerational mobility as measures of equal opportunities. In contrast to their work, we directly estimate inequality of opportunity statistics. Therefore, our focus is not on downward bias that follows from focusing on one circumstance only (i.e. parental income) but on balancing both downward and upward bias if the set of available circumstances is large in relation to a given sample size.

ity of opportunity (Van de gaer, 1993; Checchi and Peragine, 2010). They are the most widely applied formulations in the empirical literature.[4]

According to the ex-ante utilitarian view, the value of a type's opportunity set is pinned down by the expected value of its outcomes, $\mathbb{E}[y|\omega]$. Thus, the distribution of opportunities in a population can be expressed by the following counterfactual distribution y^C:

$$y^C = (y_1^C, ..., y_i^C, ..., y_N^C) = (\mathbb{E}[y_1|\omega_1], ..., \mathbb{E}[y_i|\omega_i], ..., \mathbb{E}[y_N|\omega_N]). \tag{2}$$

From this distribution one can construct ex-ante utilitarian measures of inequality of opportunity by choosing any functional $I()$ that satisfies the following two properties:

1. $I(y^C)$ decreases (increases) through transfers from i to j if i is from a circumstance type with a higher (lower) expected value of outcomes than the recipient j.
2. $I(y^C)$ remains unaffected by transfers from i to j if they are members of the same type.

In most empirical applications $I()$ represents an inequality index satisfying the standard properties of anonymity, the principle of transfers, population replication, and scale invariance (Cowell, 2016).[5] Examples of the latter are the Gini index or any member of the generalized entropy class. Note that the choice of $I()$ is normative in itself as it specifies the extent of inequality aversion at different points of the counterfactual distribution y^C. For example, the mean logarithmic deviation (MLD) values compensating transfers to the most disadvantaged types more than the Gini index. In this work we are agnostic about the normatively correct choice of $I()$. While we present our main results in terms of the Gini index, we provide robustness checks based on other inequality indexes in Supplementary Material S.6.

[4]The use of machine learning methods is not restricted to ex-ante utilitarian formulations and can be easily extended to alternative measures of inequality of opportunity.

[5]The β coefficient from intergenerational mobility regressions can also be interpreted as an ex-ante utilitarian measure of inequality of opportunity. In the intergenerational mobility framework, $\beta = \frac{E(y_{ic}|y_{ip})}{y_{ip}}$, where y_{ip} represents parental income as the sole circumstance. Hence, the functional applied to the distribution of conditional expectations can be written as $I() = \frac{1}{y_{ip}}$. Note that β decreases (increases) through transfers from children from advantaged (disadvantaged) backgrounds to children from less (more) advantaged backgrounds. However, β remains unaffected by transfers between children from parental households with equal y_{ip}.

Estimation. Given the measurement decisions described above, we require an estimate of the conditional distribution y^C. The data generating process (DGP) described in equation (1) can be rewritten as follows:

$$y = d(\omega, \theta) = f(\omega) + \epsilon = \mathbb{E}(y|\omega) + \epsilon. \tag{3}$$

$\mathbb{E}(y|\omega)$ captures unfair variation due to observed circumstances. The iid error term ϵ captures both fair (individual effort) and unfair (unobserved circumstances) determinants of individual outcomes; hence resulting measures of inequality of opportunity have a lower bound interpretation.

Estimating y^C is a prediction task in which the researcher tries to answer the following question: What outcome y_i do we expect for an individual that faces circumstances ω_i? The precise form of $f()$ is a priori unknown. In the vast majority of empirical applications, researchers address this lack of knowledge by invoking strong functional form assumptions. For example, they perform a log-linear regression of the outcome of interest on the set of observed circumstances and construct an estimate for y^C from the predicted values:

$$\ln(y_i) = \beta_0 + \sum_{p=1}^{P} \beta_p \omega_i^p + \epsilon_i, \tag{4}$$

$$\hat{y}_i^C = \exp\left[\beta_0 + \sum_{p=1}^{P} \hat{\beta}_p \omega_i^p\right]. \tag{5}$$

The literature refers to this estimation procedure as the *parametric approach* (Bourguignon et al., 2007; Ferreira and Gignoux, 2011).[6]

According to another procedure, the researcher partitions the sample into mutually exclusive types based on the realizations of all circumstances under consideration. An estimate for y^C is then constructed from average incomes within types:

$$\hat{y}_i^C = \mu_{m(i)} = \frac{1}{N_m} \sum_{j=1}^{N_m} y_j, \ \forall j \in t_m, \ \forall t_m \in \mathcal{T}. \tag{6}$$

[6]The logarithmic transformation is not innocuous as the marginal impact of circumstances on incomes may differ from their impact on log-incomes. Therefore, the predicted outcome should be obtained by applying the correction suggested in Blackburn (2007). This correction, however, is rarely implemented in empirical applications.

The literature refers to this estimation procedure as the *non-parametric approach* (Checchi and Peragine, 2010).

Both approaches face empirical challenges that are typically resolved by discretionary decisions of the researcher. For example, the parametric approach assumes a log-linear impact of all circumstances and therefore neglects the existence of interdependencies between circumstances and other non-linearities. To alleviate this shortcoming the researcher may integrate interaction terms and higher order polynomials into equation (4). However, such extensions remain at her discretion. Reversely, the non-parametric approach does not restrict the interdependent impact of circumstances. However, if the data is rich enough in information on circumstances, the researcher may be forced to reduce the observed circumstance vector to obtain statistically meaningful estimates of the relevant parameters.[7] The necessary process of restricting the circumstance vector again remains at the researcher's discretion.

The previous discussion illustrates that common approaches leave the researcher to her own devices when selecting the best model for estimating the distribution y^C. In this paper, we provide an automated solution to this problem. Similarly, Li Donni et al. (2015) propose the use of latent class modeling to obtain type partitions that allow for estimates of y^C according to the non-parametric procedure outlined in equation (6). In their approach, observable circumstances are considered indicators of membership in an unobservable latent type. For each possible number of latent types individuals are assigned to types so as to minimize the within-type correlation of observable circumstances. Then the optimal number of types, M^*, is selected by minimizing an appropriate model selection criterion such as Schwarz's Bayesian Information Criterion (BIC). The latent class approach therefore partly solves the issue of arbitrary model selection. However, it has important drawbacks. First, it cannot solve the problem of model selection once the potential number of types exceeds the available degrees of freedom. In such cases, the latent class approach replicates the limitations of parametric and non-parametric approaches: the researcher must pre-select circumstances and their subpartition. Second, latent classes are obtained by minimizing the within-type correlation of circumstances while ignoring the correlation of circumstance variables with the outcome variable. As a consequence, they are

[7] Assume the researcher observes ten circumstance variables with three expressions each—a quantity easily observed in many data sets. The non-parametric approach would require the estimation of $3^{10} = 59,049$ group means.

likely to underfit the data leading to downward biased estimates of inequality of opportunity (Lanza et al., 2013).

In the following section, we will discuss how regression trees and forests address the outlined shortcomings of existing estimation approaches.

3 ESTIMATING INEQUALITY OF OPPORTUNITY FROM REGRESSION TREES AND FORESTS

Regression trees and forests belong to the class of supervised learning methods that were developed to make out-of-sample predictions of a dependent variable based on a number of observable predictors. As we will outline in the following, they can be straightforwardly applied to inequality of opportunity estimations and solve the issue of model selection.

First, we will introduce conditional inference regression trees. By providing predictions based on identifiable groups, they closely connect to Roemer's theoretical formulation of inequality of opportunity.[8] Second, we will introduce conditional inference forests, which are—loosely speaking—a collection of many conditional inference trees. While forests do not have the intuitive appeal of regression trees, they perform better in terms of out-of-sample prediction accuracy and hence provide better estimates of the counterfactual distribution y^C.

3.1 Conditional Inference Trees

Trees obtain predictions for outcome y as a function of input variables $x = (x^1, ..., x^k)$. They use the sample $\mathcal{S} = \{(y_i, x_i)\}_{i=1}^S$ to divide the population into non-overlapping groups, $\mathcal{G} = \{g_1, ..., g_m, ..., g_M\}$, where each group g_m is homogeneous in the expression of some input variables. These groups are called *terminal nodes* or *leafs*. The conditional expectation for observation i is estimated from the mean outcome $\hat{\mu}_m$ of the group g_m to which i is assigned. Hence, in addition to the observed outcome vector $y = (y_1, ..., y_i, ..., y_N)$ one obtains a vector of predicted

[8] Furthermore, their simple graphical illustration may be an instructive tool for comparisons of opportunity structures in different societies.

values $\hat{y} = (\hat{f}(x_1), ..., \hat{f}(x_i), ..., \hat{f}(x_N))$, where

$$\hat{f}(x_i) = \hat{\mu}_{m(i)} = \frac{1}{N_m} \sum_{j \in g_m} y_j. \tag{7}$$

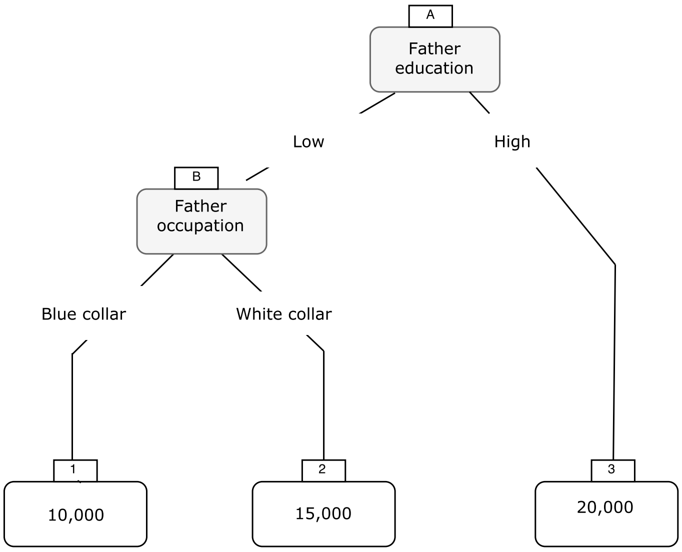
The mapping from regression trees to equality of opportunity estimation is straightforward. If the input variables $x = (x^1, ..., x^k)$ are circumstances only, each resulting group $g_m \in \mathcal{G}$ can be interpreted as a circumstance type $t_m \in \mathcal{T}$. Furthermore, $\hat{y}$ is analogous to an estimate of the counterfactual distribution y^C that underpins the construction of ex-ante utilitarian measures of inequality of opportunity.

Tree Construction. Regression trees partition the sample into M types by *recursive binary splitting*. Recursive binary splitting starts by dividing the full sample into two distinct groups according to the value they take in one input variable $\omega^p \in \Omega$. If ω^p is a continuous or ordered variable, then $i \in t_l$ if $\omega_i^p < \tilde{\omega}^p$ and $i \in t_m$ if $\omega_i^p \geq \tilde{\omega}^p$, where $\tilde{\omega}^p$ is a splitting value chosen by the algorithm. If ω^p is a categorical variable then the categories can be split into any two arbitrary groups. The process is continued such that one of the two groups is divided into further subgroups (potentially based on another $\omega^q \in \Omega$), and so on. Graphically, this division into groups can be presented like an upside-down tree (Figure 1).

The exact manner in which the split is conducted depends on the type of regression tree that is used. In this paper, we follow the conditional inference methodology proposed by Hothorn et al. (2006). Conditional inference trees are grown by a series of permutation tests according to the following 4-step procedure:

0. Choose a significance level α^*.
1. Test the null hypothesis of density function independence: $H_0^{\omega^p} : D(y|\omega^p) = D(y)$, for all $\omega^p \in \Omega$, and obtain a p-value associated with each test, p^{ω^p}.
   - $\Rightarrow$ Adjust the p-values for multiple hypothesis testing, such that $p_{adj.}^{\omega^p} = 1 - (1 - p^{\omega^p})^P$ (Bonferroni Correction).

FIGURE 1 – Exemplary Tree Representation



Note: Artificial example of a regression tree. Gray boxes indicate splitting points; white boxes indicate terminal nodes. The values inside terminal nodes show estimates for the conditional expectation y^C.

2. Select the variable ω^* with the lowest p-value, i.e.

$$\omega^* = \underset{\omega^p}{\operatorname{argmin}}\{p_{adj.}^{\omega^p} : \omega^p \in \Omega,\ p = 1, ..., P\}.$$

   $\Rightarrow$ If $p_{adj.}^{\omega^*} > \alpha^*$: Exit the algorithm.

   $\Rightarrow$ If $p_{adj.}^{\omega^*} \leq \alpha^*$: Continue, and select ω^* as the splitting variable.

3. Test the null hypothesis of density function independence between the subsamples for each possible binary partition splitting point s based on ω^*, and obtain a p-value associated with each test, $p^{\omega_s^*}$.

   $\Rightarrow$ Split the sample based on ω^*, by choosing the splitting point s that yields the lowest p-value, i.e. $\tilde{\omega}^* = \underset{\omega_s^*}{\operatorname{argmin}}\{p^{\omega_s^*} : \omega_s^* \in \Omega\}$.

4. Repeat steps 1.–3. for each of the resulting subsamples.

In words, conditional inference trees start by a series of univariate hypothesis tests. The circumstance that is most related to the outcome is chosen as the potential splitting variable. If the dependence between the outcome and the splitting variable is sufficiently strong, then a split is made. If not, no split is made. Whenever a circumstance can be split in several ways, the sample is split into two subsamples such that the dependence with the outcome variable is maximized. This procedure is repeated in each of the two subsamples until no circumstance in any subsample is sufficiently related to the outcome variable. Note that the depth of the resulting opportunity tree hinges on the level of α^*. The less stringent the α^*-requirement, the more we allow for false positives, i.e. the more splits will be detected as significant and the deeper the tree will be grown. In our empirical application we fix $\alpha^* = 0.01$, which is in line with the disciplinary convention for hypothesis tests. To illustrate the robustness of this choice we show comparisons to setting $\alpha^* = 0.05$ and choosing α^* through cross-validation in Supplementary Figure S.1.

A particular advantage of trees is that they avoid list-wise deletion of observations by implementing surrogate splits. In case of missing data, the algorithm searches for an alternative splitting point that mimicks the sample partition based on $\tilde{\omega}^*$ to the greatest extent. All observations that lack information on $\tilde{\omega}^*$ are then allocated to subbranches based on this surrogate splitting point.

3.2 *Conditional Inference Forests*

Regression trees provide a simple and standardized way of dividing the population into types. Therefore, they solve the model selection problem outlined in section 2. However, trees suffer from three shortcomings: first, the structure of trees—and therefore the estimate of y^C—is fairly sensitive to alternations in data samples. This issue is particularly pronounced if there are various circumstances that are close competitors for defining the first splits (Friedman et al., 2009). Second, trees assume a non-linear data generating process that imposes interactions while ruling out the linear influence of circumstances. Third, trees make inefficient use of data since some of the circumstances $\omega^p \in \Omega$ are not used for the construction of the tree. However, circumstances may possess informational content that can increase predictive power even if they are not significantly associated with y at level α^*. This becomes an issue if two or more important circumstances are highly correlated. Once a split is made using either of the two, it is unlikely that the other contains enough information to cause another split. Conditional inference forests address all of these shortcomings (Breiman, 2001; Biau and Scornet, 2016).

Forest Construction. Random forests create many trees and average over all of these when making predictions. Trees are constructed according to the same 4-step procedure outlined in the previous subsection. However, two tweaks are made. First, given the sample $\mathcal{S} = \{(y_i, \omega_i)\}_{i=1}^{S}$ each tree is estimated on a random subsample $\mathcal{S}' \subset \mathcal{S}$. In our application, we randomly select approximately 60% of the observations for each tree, and estimate B^* such trees in total. Second, only a random subset of circumstances of cardinality $\bar{P}^*$ is allowed to be used at each splitting point. Together these two tweaks remedy the shortcomings of single conditional inference trees. First, averaging over B^* predictions cushions variance in the estimates of y^C and smooths the non-linear impact of circumstance characteristics. Second, using subsets of all circumstance variables increases the likelihood that all observed circumstances with informational content will be identified as splitting variable ω^* at some point.

Predictions are formed as follows:

$$\hat{f}(\omega; \alpha^*, \bar{P}^*, B^*) = \frac{1}{B^*} \sum_{b=1}^{B^*} \hat{f}^b(\omega; \alpha^*, \bar{P}^*). \tag{8}$$

Equation (8) illustrates that individual predictions are a function of α^*—the significance level governing the implementation of splits, $\bar{P}^*$ —the number of circumstances to be considered at each splitting point, and B^*—the number of subsamples drawn from the data. In our empirical illustration we fix $B^* = 200$ and determine α^* and $\bar{P}^*$ by minimizing the *out-of-bag* error (MSE^{OBB}). Details on these choices and empirical procedures are disclosed in Supplementary Material S.1.

4 EMPIRICAL APPLICATION

In this section we illustrate the machine learning approach using harmonized survey data from 31 European countries. We compare the results from trees and forests with results from the prevalent estimation approaches in the extant literature; namely parametric, non-parametric and latent class models. Comparisons are made along two dimensions.

First, we evaluate the different estimation approaches by comparing their out-of-sample mean squared error (MSE^{Test}). MSE^{Test} is a standard statistic to evaluate the prediction quality of estimation models.[9] To calculate MSE^{Test}, we follow the machine learning practice of splitting our sample into a *training set* with $i^{-H} \in \{1, ..., N^{-H}\}$ and a *test set* with $i^{H} \in \{1, ..., N^{H}\}$. For each sample, we choose $N^{-H} = \frac{2}{3}N$ and $N^{H} = \frac{1}{3}N$.[10] We fit our models on the training set and compare their performance on the test set according to the following procedure:

1. Run the model on the training data (for the specific estimation procedures, see section 3.1 for trees and forests, and section 4.2 for our benchmark methods).
2. Store the prediction function $\hat{f}^{-H}()$.
3. Calculate the mean squared error in the test set:
   $\text{MSE}^{\text{Test}} = \frac{1}{N^H} \sum_{i \in H} [y_i - \hat{f}^{-H}(\omega_i)]^2$.

Second, we evaluate the different approaches by comparing inequality of opportunity estimates. To this end, we run the models on all data for a country, and apply the resulting prediction functions $\hat{f}()$ to obtain $\hat{y}^C$. Estimates of inequality of opportunity are derived by summarizing $\hat{y}^C$ with the Gini index. Estimates for alternative inequality indexes are presented in Supplementary Material S.6.

4.1 Data

We base our empirical illustration on the 2011 wave of the European Union Statistics on Income and Living Conditions (EU-SILC). EU-SILC provides harmonized survey data with respect to

[9] Minimizing MSE^{Test} is equivalent to trading-off upward and downward biases of inequality of opportunity estimates in a given data environment: The more parsimonious the model, the higher the prediction bias (underfitting) and the stronger the downward bias in inequality of opportunity estimates. The more complex the model, the higher the prediction variance (overfitting) and the stronger the upward bias of inequality of opportunity estimates. We provide a thorough illustration of this mapping in Supplementary Material S.2.

[10] Note that the size of the training set for each country is constant regardless of the estimation method. Hence, any cross-method differences in prediction accuracy are not driven by differences in sample size.

income, poverty, and living conditions. It is the official reference source for comparative statistics on income distribution and social inclusion in the European Union. In its 2011 wave EU-SILC covers a cross-section of 31 European countries. For each country, it contains a random sample of all resident private households. Data is collected by national statistical agencies following common variable definitions and data collection procedures. We use the 2011 wave since it contains an ad-hoc module about the intergenerational transmission of (dis)advantages. This module allows us to construct finely-grained circumstance type partitions. Observed circumstances Ω and their respective expressions are listed in Table 1. We include all variables of EU-SILC containing information about the respondent's characteristics at birth and their living conditions during childhood. Descriptive statistics of circumstance variables are reported in Supplementary Material S.5.[11]

The unit of observation is the individual and the outcome of interest is equivalized disposable household income. We obtain the latter by dividing household disposable income with the square root of household size. Reported incomes refer to the year preceding the survey wave, i.e. 2010 in the case of our empirical application. In line with the literature we focus on equivalized household income as it provides the closest income analogue to consumption possibilities and general economic well-being. Inequality statistics tend to be heavily influenced by outliers (Cowell and Victoria-Feser, 1996); therefore we adopt a standard winsorization method according to which we set all non-positive incomes to 1 and scale back all incomes exceeding the 99.5th percentile of the country-specific income distribution to this lower threshold. Our analysis is focused on the working age population. Therefore, we restrict the sample to respondents aged between 30 and 60. To assure the representativeness of our inequality of opportunity estimates we use individual cross-sectional weights when calculating $I(\hat{y}^C)$.

Table 2 shows considerable heterogeneity in income distributions across Europe. While the average households in Norway and Switzerland obtained incomes above €40,000 in 2010, the average household income in Romania, Bulgaria and Lithuania did not exceed the €5,000 mark. Lowest inequality prevails in Norway, Iceland and Denmark, all of which have Gini coeffi-

[11] In contrast to some existing work we do not consider age as a circumstance (among others Checchi et al., 2016). This choice is motivated by the fact that cross-sectional income disparities across age groups even out across the life-cycle of individuals. In Supplementary Material S.7, we provide robustness analyses based on income distributions that are residualized from variation across age groups. Our conclusions remain unaffected.

TABLE 1 – List of Circumstances

1. Respondent's sex:
   - Male
   - Female
2. Respondent's country of birth:
   - Respondent's present country of residence
   - European country
   - Non-European country
3. Presence of parents at home*:
   - Both present
   - Only mother
   - Only father
   - Without parents
   - Lived in a private household without any parent
4. Number of adults (aged 18 or more) in respondent's household*
5. Number of working adults (aged 18 or more) in respondent's household*
6. Number of children (under 18) in respondent's household*
7. Father's/mother's country of birth and citizenship:
   - Born/citizen of the respondent's present country of residence
   - Born/citizen of another EU-27 country
   - Born/citizen of another European country
   - Born/citizen of a country outside Europe
8. Father's/mother's education (based on the International Standard Classification of Education 1997 [ISCED-97])*:
   - Unknown father/mother
   - Illiterate
   - Low (0-2 ISCED-97)
   - Medium (3-4 ISCED-97)
   - High (5-6 ISCED-97)
9. Father's/mother's occupational status*:
   - Unknown or dead father/mother
   - Employed
   - Self-employed
   - Unemployed
   - Retired
   - House worker
   - Other inactive
10. Father's/mother's main occupation (based on the International Standard Classification of Occupations, published by the International Labour Office [ISCO-08])*:
    - Managers (I-01)
    - Professionals (I-02)
    - Technicians (I-03)
    - Clerical support workers (I-04)
    - Service and sales workers (I-05 and 10)
    - Skilled agricultural, forestry and fishery workers (I-06)
    - Craft and related trades workers (I-07)
    - Plant and machine operators, and assemblers (I-08)
    - Elementary occupations (I-09)
    - Armed forces occupations (I-00)
    - Father/mother did not work, was unknown or was dead
11. Managerial position of father/mother*:
    - Supervisory
    - Non-supervisory
12. Tenancy status of the house in which the respondent was living*:
    - Owned
    - Not owned

Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This table lists the circumstance variables available in EU-SILC 2011. Questions marked with * refer to the time when the respondent was 14 years old. Item 7 (11) is missing for Slovenia (Finland).

cients below 0.230. At the other end of the spectrum we find Latvia and Lithuania with Gini coefficients above 0.340.

4.2 Benchmark Methods

We compare trees and forests to three benchmark estimation methods from the extant literature.

TABLE 2 – Summary Statistics

		Equivalized Disposable Household Income in €		
Country	N	μ	σ	Gini
Austria	6,220	25,538	13,408	0.267
Belgium	6,011	23,314	10,769	0.247
Bulgaria	7,146	3,698	2,457	0.331
Croatia	6,945	6,631	3,764	0.306
Cyprus	4,589	21,074	11,554	0.278
Czech Republic	8,711	9,036	4,610	0.253
Denmark	5,795	32,471	14,422	0.227
Estonia	5,338	6,924	4,364	0.330
France	11,078	24,320	14,695	0.287
Germany	12,683	22,862	12,468	0.284
Greece	6,184	13,184	8,887	0.334
Hungary	13,330	5,305	2,830	0.275
Iceland	3,682	21,562	9,290	0.221
Ireland	4,318	24,882	14,078	0.295
Italy	21,070	18,774	11,348	0.314
Latvia	6,423	5,339	3,751	0.362
Lithuania	5,403	4,774	3,116	0.344
Luxembourg	6,765	37,948	19,412	0.270
Malta	4,701	13,058	6,758	0.272
Netherlands	11,411	24,322	11,452	0.243
Norway	5,026	42,265	16,679	0.206
Poland	15,545	6,087	3,837	0.316
Portugal	5,899	10,796	7,354	0.333
Romania	7,820	2,527	1,612	0.336
Slovakia	6,779	7,309	3,509	0.256
Slovenia	13,183	13,373	5,896	0.234
Spain	15,481	17,088	10,684	0.328
Sweden	6,599	25,098	11,157	0.237
Switzerland	7,583	42,844	23,877	0.278
United Kingdom	7,391	22,768	15,164	0.319

Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This table provides summary statistics by country. N indicates the total number of observations. The last three columns summarize the distribution of equivalized disposable household income: mean (μ), standard deviation (σ), and Gini coefficient.

First, we draw on the parametric approach as proposed by Bourguignon et al. (2007) and Ferreira and Gignoux (2011). In line with equation (4), estimates are obtained by a Mincerian regression of log income on the following circumstances: educational attainment of mother and father (5 categories each), father's occupation (11 categories), area of birth (3 categories), and tenancy status of the household at age 14 (2 categories). The prediction model includes 22 parameters.

Second, we draw on the non-parametric approach as proposed by Checchi and Peragine (2010). In line with equation (6), non-parametric estimates are obtained by calculating average outcomes in non-overlapping circumstance types. Types are homogeneous with respect to educational attainment of the highest educated parent (5 categories), fathers' occupation (4 categories), and migration status (2 categories).[12] The prediction model includes 40 parameters.

[12] To minimize the frequency of sparsely populated types we divert from the occupational list given in Table 1 by

Third, we draw on the latent class approach as proposed by Li Donni et al. (2015). We use the union of circumstances used in the parametric and non-parametric approach from which the algorithm infers the appropriate number of unobserved types in the data by minimizing BIC.

Do these specification choices serve for a fair assessment of these benchmark methods? As outlined in section 2, model specification in the (non-)parametric approach is a discretionary choice of the researcher; therefore there are many different specifications that could be used for the benchmarking. To make the comparison non-arbitrary, we anchor our comparison on model specifications of existing studies. The specification of the parametric approach is inspired by Palomino et al. (2019). We divert from their specification by excluding gender (due to our focus on disposable household income) and retrospective information on the financial situation during childhood (due to potential recall bias) from the list of circumstances. In comparison, our prediction model (22 parameters) is more parsimonious than the model in Palomino et al. (2019, 24 parameters). The specification of the non-parametric approach is inspired by Checchi et al. (2016). We divert from their specification by excluding gender (due to our focus on disposable household income) and age (due to its interpretation as a proxy for life-cycle effects) from the list of circumstances. In comparison, our prediction model (40 parameters) is more parsimonious than the model in Checchi et al. (2016, 96 parameters). As outlined in section 2, model specification in latent class analysis is data driven. We therefore do not need to specify the model itself but commit to a model selection criterion. We anchor our comparison on the study of Li Donni et al. (2015) who use BIC to select the number of latent classes to be estimated.

4.3 Simulation

We begin our analysis with a simulation exercise. The simulation allows us to assess the properties of different estimation approaches while maintaining control over the true DGP. As a consequence we can i) assess the prediction accuracy by decomposing MSE^{Test} into its variance and bias components, and ii) assess the resulting bias in inequality of opportunity estimates.

re-coding occupations into the following categories: high-skilled non-manual (I-01–I-03), low-skilled non-manual (I-04–I-05 and I-10), skilled manual and elementary occupation (I-06–I-09), and unemployed/unknown/dead.

In general, simulation results are sensitive to assumptions about the true DGP and sample sizes. To make the simulation relevant to the context of our empirical analysis, we choose DGPs that are anchored in the benchmark methods presented in section 4.3 and choose sample sizes to broadly cover the range of country samples in EU-SILC. We note that our additional simulation choices are conservative. First, we construct a simulation sample without missing data points. As a consequence—and in contrast to actual empirical applications—parametric and non-parametric approaches do not suffer from data reductions through list-wise deletion. Second, we restrict circumstances used by trees and forests to the union of circumstances used in the (non-)parametric approach. As a consequence—and in contrast to actual empirical applications—we deprive data-driven approaches from the advantage of using all available circumstance information in the data.

We impose three DGPs that are summarized in Table 3. The *parametric DGP* and *non-parametric DGP* correspond to the estimation models outlined in section 4.2. They present a challenging test for data-driven estimation methods since the latter have to compete against fixed specifications (parametric, non-parametric) that correspond to the ground truth. In addition, we specify a *mixed DGP* that integrates features of both the parametric and the non-parametric DGP. This is a more realistic scenario since it is plausible to assume that researchers devise fixed specifications without prior knowledge of the true DGP. We estimate all three models on the full sample of EU-SILC while list-wise deleting observations with missing information ($N = 197,565$). In turn, we retain the predictions from these estimations and add a disturbance term with $\mathcal{N}(0, 2000)$.[13] Thus, we obtain three variables that define the distribution of income for the purpose of this simulation.

Next we specify five sample sizes that broadly cover the range of effective country sample sizes observed in EU-SILC (see Supplementary Table S.1): $N \in \{1,000; 2,000; 4,000; 8,000; 16,000\}$. For each sample size, we draw one test set of size $N^H = 1/3N$ and 50 training sets of size $N^{-H} = 2/3N$. Thus, for each observation in the test sets we obtain 50 predictions per combination of DGP and estimation approach. Based on these predictions we calculate two statistics: i) expected MSE^{Test} to assess out-of-sample prediction accuracy (James et al., 2013), and ii) the expected absolute difference between inequality of opportunity estimates and the true level of

[13]We choose a variance term small enough such that $y \in \mathbb{R}_{++}$.

TABLE 3 – Summary of Data Generating Processes

	Parametric	Non-Parametric	Mixed
Outcome	ln(y)	y	ln(y)
Parameters	22	40	18
Circumstances	Education Father Education Mother Occupation Father Birth Area Tenancy Status	Education Parents Occupation Father Mig. Background	Education Parents Occupation Father Mig. Background Tenancy Status
Non-Linearity	None	Full Interaction	All Circ. w/ Mig. Background (2 Levels)
ϵ	$\mathcal{N}(0, 2000)$	$\mathcal{N}(0, 2000)$	$\mathcal{N}(0, 2000)$

inequality of opportunity.

Figure 2 displays the results of our simulation. In its lower part, each panel describes expected MSE^{Test} per combination of DGP, estimation approach and sample size. Since we know the true DGP we can decompose MSE^{Test} into variance and expected bias.[14] In its upper part, each panel describes the corresponding absolute bias in inequality of opportunity estimates on an inverse scale. The absolute bias is calculated as the expected absolute difference between inequality of opportunity estimates and the true level of inequality of opportunity as a percentage share of the latter.

The simulation results are in line with statistical theory. First, if fixed estimation approaches (parametric, non-parametric) invoke the true DGP, expected bias is zero and MSE^{Test} is driven by its variance component only. Second, with increasing N, the bias component of MSE^{Test} remains constant for fixed specifications (parametric, non-parametric) and decreases for data-driven approaches (LCA, trees, forests). Third, with increasing N the variance component of MSE^{Test} decreases for all combinations of DGPs and estimation approaches. Fourth, forests tend to have lower variance than trees—in our simulation this is true in 93% of all cases.

Furthermore, Figure 2 illustrates that the size of MSE^{Test}, and therefore bias in inequality of opportunity estimates, varies with sample size for all estimation methods. However, the sources of this bias vary across estimation methods. For example, forests incur downward bias in

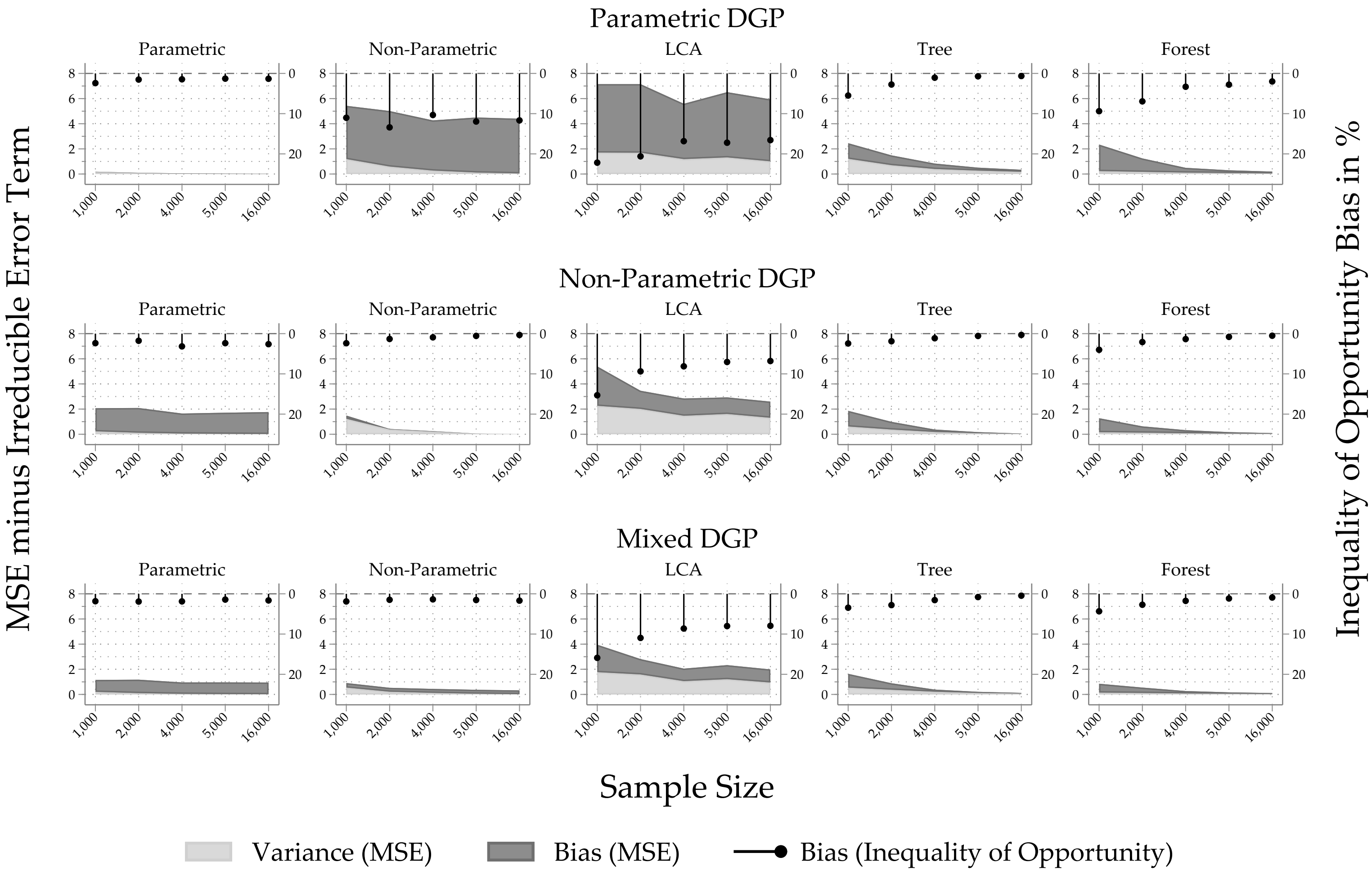
[14]See also Supplementary Material S.2 for an illustration of the variance-bias decomposition. The irreducible error term is uninformative for differences in MSE^{Test} because $\text{Var}(\epsilon) = 2,000^2$ is constant across specifications. Therefore, we only present evidence on the variance and the bias component of MSE^{Test}.

small samples as their algorithm prevents the detection of relevant splits. To the contrary, fixed specifications incur upward bias in small samples as the underlying parameters are noisily estimated (although unbiased in expectation). The crucial question is whether this sensitivity is larger for ML than for traditional econometric methods which is a case-specific and empirical question.

Under the reasonable assumption that researchers do not know the true DGP, forests clearly dominate all other estimation approaches in terms of expected MSE^{Test}. This result holds both in comparison to fixed estimation approaches (parametric, non-parametric) and in comparison to LCA as an alternative data-driven estimation approach. The results for trees are also persuasive, however, they have a weaker performance than forests when samples are small. Even in the unlikely case that researchers were to specify (non-)parametric models correctly, trees and forests quickly converge to the test error of the fixed model that invokes the true DGP. The simulation results furthermore highlight the close correspondence between MSE^{Test} and expected bias in inequality of opportunity estimates: the higher MSE^{Test}, the stronger inequality of opportunity estimates diverge from the ground truth.

In summary, the simulation results support the use of regression trees and forests. They flexibly approximate the true DGP. Thereby, they outperform fixed estimation approaches (parametric, non-parametric) and alternative data-driven estimation approaches (LCA) in terms of the expected MSE^{Test} which itself is tightly linked to expected bias in inequality of opportunity estimates.

FIGURE 2 – Simulation Results



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows expected MSE^{Test} and expected bias in inequality of opportunity estimates from the simulation exercise. Each row corresponds to one data generating process (see Table 3). Each column corresponds to one estimation method (see sections 3.1, 3.2, 4.2). We multiply MSE^{Test} by 1×10^{-6} and deduct the irreducible error term. Inequality of opportunity is measured by the Gini coefficient of the counterfactual distribution $\hat{y}^C$. We measure expected bias in inequality of opportunity by the average absolute difference between inequality of opportunity estimates and the true level of inequality of opportunity as specified by the data generating process.

4.4 *Cross-Country Comparison*

We now turn to a cross-country comparison based on actual data. First, we calculate MSE^{Test} to assess the prediction accuracy of different estimation approaches. Second, we calculate inequality of opportunity estimates. In contrast to the simulation exercise, we do not know the true DGP and we cannot assess bias in inequality of opportunity estimates by comparison to the ground truth. Therefore, we assess bias in inequality of opportunity estimates by comparing estimation approaches against the method with the highest prediction accuracy, i.e. the method yielding the lowest MSE^{Test}.

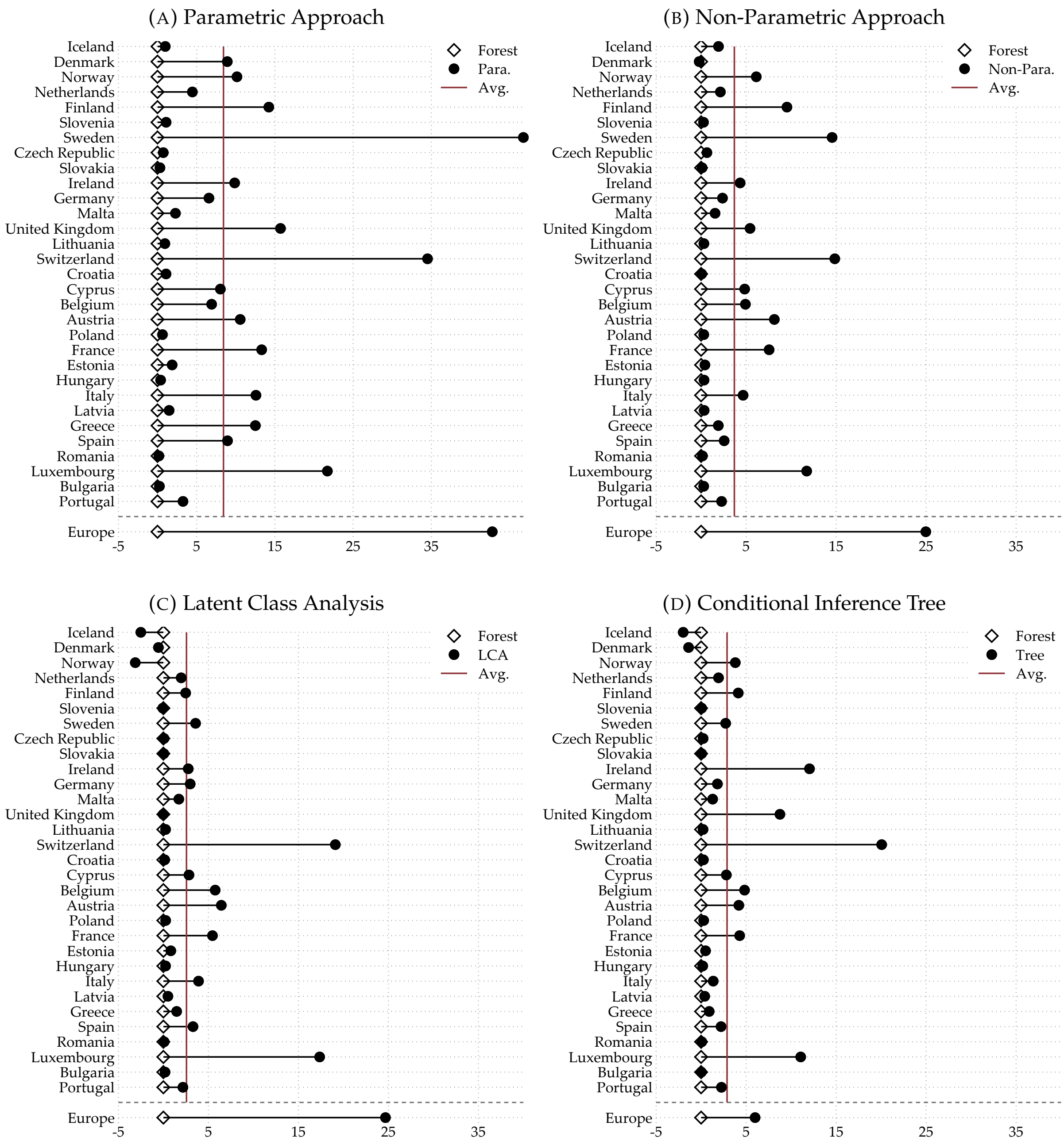
Prediction Accuracy. Figure 3 compares MSE^{Test} across countries and estimation approaches. For each method, MSE^{Test} is presented in differences relative to random forests. By differencing across methods, we provide a close analogue to the simulation exercise in section 4.3: we omit the irreducible error term from the comparison, and relative MSE^{Test} is driven by variance and bias components, only. For better visual clarity, we again scale MSE^{Test} by 1×10^{-6}. Relative $\text{MSE}^{\text{Test}} > 0$ indicates poorer prediction accuracy in comparison to random forests.

Random forests outperform all other methods in terms of prediction accuracy. On average, the parametric approach yields test errors that exceed random forests by 8.4 (7.8%) (Figure 3, Panel [a]). Somewhat smaller average shortfalls in prediction accuracy are observed for non-parametric (Figure 3, Panel [b]) and latent class models (Figure 3, Panel [c]). Averages across countries, however, mask considerable heterogeneity. For example, the relative test error of parametric estimates for Eastern European countries like Slovenia or Czech Republic are close to zero. To the contrary, relative test errors of parametric estimates for Sweden, Luxembourg, and Switzerland diverge significantly from the forest benchmark.

Conditional inference trees are closest to the test error rate of forests ($\text{MSE}^{\text{Test}} = 2.8$ [2.2%]). Yet, they also fall short of the performance of forests due to higher variance, imposing non-linearity, and omitting less relevant circumstances (see section 3.2).

We conclude that among all considered methods, conditional inference forests deliver the high-

FIGURE 3 – Comparison of MSE$^{\text{Test}}$ by Method



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows differences of MSE$^{\text{Test}}$ from different estimation approaches relative to random forests. For all methods, we multiply MSE$^{\text{Test}}$ by 1×10^{-6}. Values $> (<) 0$ indicate worse (better) out-of-sample prediction accuracy than random forests. Vertical lines indicates unweighted cross-country averages. Point estimates and associated standard errors are listed in Supplementary Table S.1.

est out-of-sample prediction accuracy.[15] Hence, relative to random forests, other methods underutilize or overutilize the information contained in Ω which in expectation will lead to bias

[15] In Supplementary Table S.3 we show that the overwhelming majority of detected differences are statistically significant at conventional levels.

in inequality of opportunity estimates.

Inequality of Opportunity Estimates. Figure 4 displays inequality of opportunity estimates across countries and estimation approaches. In each panel we plot inequality of opportunity estimates for a particular method, as well as the associated differences to estimates from forests. We emphasize that results from forests cannot be interpreted as the truth. However, since forests yield the lowest test error among all considered methods, they provide the best *approximation of the true DGP in a given data environment*. Therefore, they are a useful benchmark to assess bias of other estimation methods.[16]

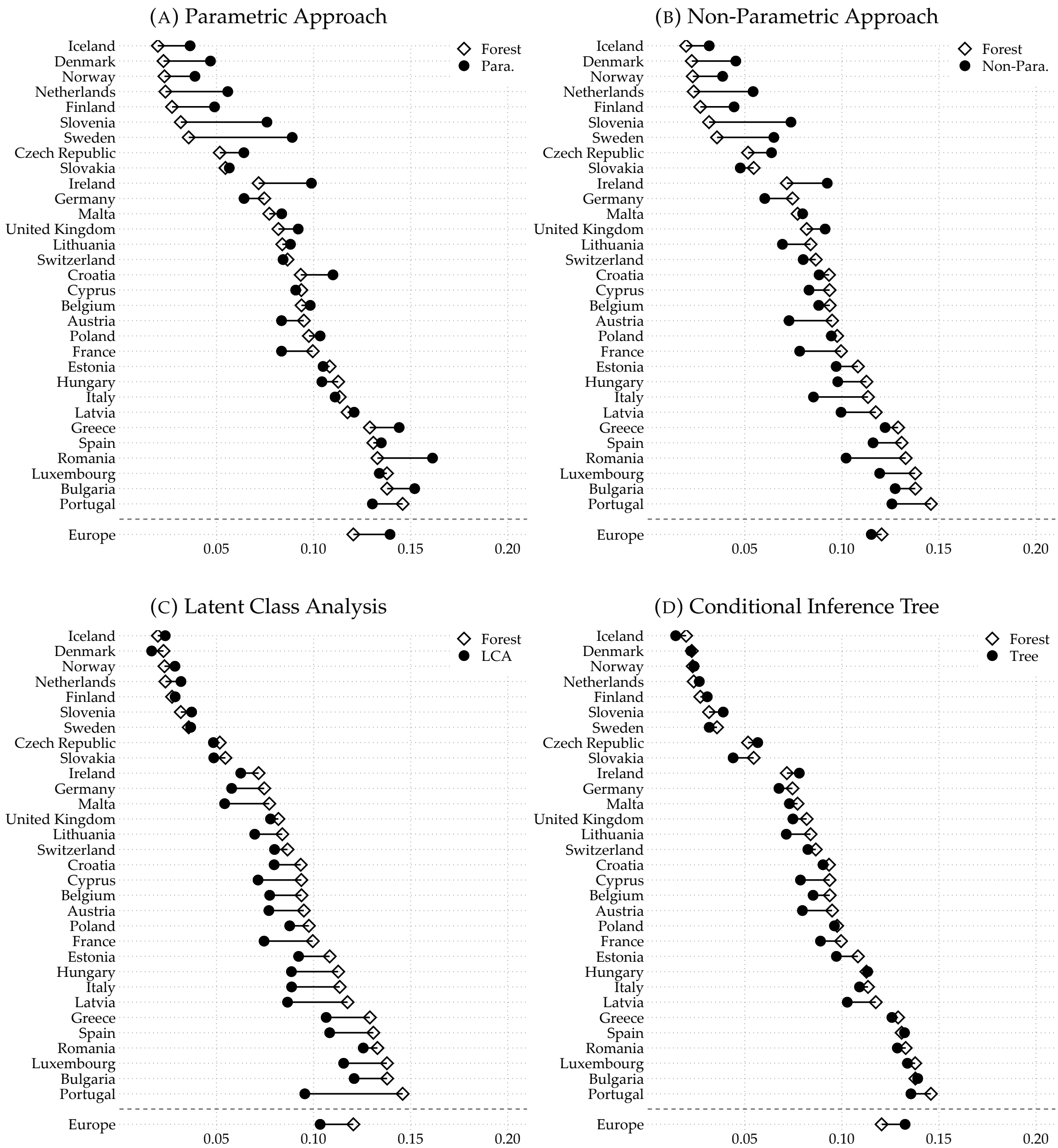
Panel (a) shows estimates from the parametric approach. In our country sample the chosen model specification for the parametric approach tends to overstate inequality of opportunity relative to forests which is the method providing the lowest expected bias in comparison to the true DGP. For 21 out of 31 countries the inequality of opportunity estimates are higher than the results from forests. Most pronounced overstatements are observed in countries that are typically considered high-opportunity societies. For example, forests classify Sweden and the Netherlands as societies offering high equality of opportunity. To the contrary, the parametric estimate would rank them at similar levels like Germany and France.

Panel (b) shows estimates from the non-parametric approach. The overall pattern is more heterogeneous than for the parametric approach. While overstatements prevail in countries that are typically considered as high-opportunity societies, there are 20 out of 31 countries for which the non-parametric estimate falls short of the forest estimate. These countries are clustered in the lower part of the equal-opportunity ranking. For example, forests classify Italy at a worse position than most countries in Europe. To the contrary, the non-parametric estimate would elevate Italy towards the mid-field close to Sweden.

We highlight that any resemblance between forests and (non-)parametric estimation approaches

[16] It is important to keep this relative interpretation of "bias" in mind. We compare method-specific estimates to the best estimate of inequality of opportunity in a given data environment. In light of all methods lacking information on unobserved circumstances, methods that are upward biased in this comparison may potentially be closer to the ground truth than our reference estimate. Such conclusions, however, are purely speculative and can neither be verified nor falsified without knowledge of the ground truth (see also Supplementary Material S.2 for a thorough explanation). Therefore, another interpretation of forests is that they provide the reliable maximum lower bound estimate of inequality of opportunity in a given data environment.

FIGURE 4 – Comparison of Inequality of Opportunity Estimates by Method



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows inequality of opportunity estimates from different estimation methods relative to forests. Inequality of opportunity is measured by the Gini coefficient of the counterfactual distribution $\hat{y}^C$. Point estimates and associated standard errors are listed in Supplementary Table S.2.

should be interpreted as a luck of the draw rather than a property inherent to the estimation approach. Under alternative plausible model specifications, estimates from the both approaches may diverge much more strongly than under the specifications adopted in this work. This property of fixed model specifications is apparent from the simulation results in section 4.3.

Panel (c) shows estimates from the latent class approach. In our country sample, LCA tends to understate inequality of opportunity relative to forests. For 25 out of 31 countries the LCA estimate falls short of its forest-based analogue. LCA chooses rather coarse type partitions. Therefore, understatements are clustered in the lower tail of the equal-opportunity ranking, i.e. in societies in which many circumstances co-produce the outcome of interest. To the contrary, in high-opportunity societies the parsimonious models chosen by LCA tend to replicate the results from forests reasonably well.

Panel (d) shows that trees and forests tend to produce similar results. The correlation between point estimates is high (0.98). In contrast to all other approaches there is no general tendency to over- or underestimate inequality of opportunity relative to forests.

Finally, detected differences between the benchmark estimation approaches and forests persist when estimating equality of opportunity in a pooled European sample.[17] For example, the parametric approach overestimates inequality of opportunity relative to forests, whereas LCA yields lower estimates than forests.

Robustness to Differences in Sample Size. Effective sample sizes differ by estimation method and country (Supplementary Table S.1). First, samples for the benchmark methods (parametric, non-parametric, LCA) are reduced as they rely on list-wise deletion in case of missing circumstance information. These reductions can be sizable and exceed 50% in 6 countries of our sample (Denmark, Iceland, Netherlands, Norway, Slovenia, Sweden). Second, even when accounting for missing information the largest country sample in EU-SILC (Italy, $N = 21,070$) is almost seven times as large as the smallest country sample (Iceland, $N = 3,682$). Therefore, we perform two robustness analyses.

First, we recompute inequality of opportunity after completing missing data through multiple imputation (Schafer, 1999).[18] As a consequence, we can compare inequality of opportunity esti-

[17] Note that we do not include country of residence as a circumstance. We acknowledge that country of residence is congruent with country of birth for most individuals. Therefore, it could be used as a proxy circumstance (Milanovic, 2015). However, our foremost concern is a methodological comparison of estimation approaches in different data environments. Therefore, we prefer to keep the set of circumstances comparable to our within-country estimates.

[18] List-wise deletion yields unbiased parameter estimates if data is missing completely at random (MCAR). Mul-

mates across methods on the same effective sample size per country. Supplementary Figure S.1 shows a decrease in inequality of opportunity estimates relative to forests for all benchmark methods (parametric, non-parametric, LCA). This result is in line with the intuition that upward biases decrease as sample sizes grow relative to the number of model parameters. To the contrary, the patterns for trees and forests remain unaffected since they handle missing values by default through surrogate splits.

Second, we recompute inequality of opportunity while reducing sample sizes across countries to the smallest common denominator. As a consequence, we can compare inequality of opportunity estimates across countries on the same effective sample size. Supplementary Figure S.2 shows that point estimates and country rankings differ for all benchmark methods (parametric, non-parametric, LCA). Furthermore, also trees show some variability as sample sizes decrease. To the contrary, point estimates and country rankings of forests are unaffected by harmonization in sample sizes across countries. On the one hand, these results highlight that the high variance of trees may lead to suboptimal results in some applications and that researchers should give preference to forest estimates where possible. On the other hand, these results bolster confidence that opportunity rankings of forests are not an artifact of cross-country variation in sample sizes.[19]

Comparison to Existing Literature. We have shown that benchmark methods from the existing literature yield markedly different estimates of inequality of opportunity relative to the method for which we expect the lowest bias. These differences are manifested in both point estimates and country rankings. Therefore, these methods may be misleading in two related dimensions. First, they may mis-classify European societies regarding their need for opportunity equalizing policy interventions. Second, researchers and policymakers in search of best practices to devise opportunity equalizing policy interventions may turn to the wrong country examples. In the following we will assess the extent to which such concerns are reflected in the

tiple imputation weakens this assumption by assuming that data is missing at random (MAR), i.e. that missing data is random conditional on observed variables.

[19] We perform a similar exercise on the pooled sample: we re-estimate our results for the pooled sample on increasingly smaller fractions of the total sample. In Supplementary Figure S.3 we again show that benchmark methods (parametric, non-parametric, LCA) and trees are sensitive to changes in sample size when fractions become small. To the contrary, forests again emerge as the method that is most robust in small samples.

extant literature on inequality of opportunity in Europe.

We proceed in two steps. First, we assess whether existing literature on inequality of opportunity in Europe is consistent, i.e. whether it yields similar opportunity rankings across European societies. If the literature were consistent, researcher discretion in model selection would be irrelevant for conclusions about inequality of opportunity in Europe. Second, we assess whether existing literature on inequality of opportunity in Europe conforms with evidence on the intergenerational income elasticity (IGE). The IGE is a commonly used proxy statistic for equality of opportunity that is based on data links across generations. The IGE provides a suitable benchmark since it can be interpreted as an ex-ante utilitarian measure of inequality of opportunity (see footnote 5) and it is often based on richer (administrative) panel data. If there was conformity, current estimation approaches would yield opportunity rankings that are strongly in line with common priors about mobility in European societies. We answer both questions by calculating correlations in opportunity rankings across i) existing studies on inequality of opportunity,[20] ii) existing consensus estimates of the IGE,[21] and iii) inequality of opportunity estimates from our preferred methods—regression trees and forests.

Panel (a) of Table 4 suggests that existing literature on inequality of opportunity in Europe is not consistent. Rank correlations as low as 0.09 indicate strong heterogeneity in country rankings. On the one hand, all studies that inform this comparison have a very high degree of harmonization in relevant dimensions: estimates were derived from the same underlying data source (EU-SILC), refer to a similar age group (25-60), and summarize counterfactual distributions $\hat{y}^C$ by the same inequality metric (mean log deviation). On the other hand, all studies specify different prediction functions to estimate inequality of opportunity. This suggests that discretionary choices with respect to model specifications may be a major force behind inconclusive evidence in the inequality of opportunity literature. We cannot fully rule out the possibility

[20]We focus on published studies estimating ex-ante measures of inequality of opportunity on the 2011 wave of EU-SILC. Further studies that do not meet both criteria include Andreoli and Fusco (2019) and Carranza (2020). Furthermore, we do not include Brzezinski (2020) since he derives estimates based on the methods proposed in this paper.

[21]We focus on IGE estimates based on actual data linkages across generations and exclude IGE estimates based on two-sample instrumental variable estimators to mitigate distortions through measurement error. Estimates are extracted from Stuhler (2018) and Carmichael et al. (2020). Jointly both studies contain the following subset of our country sample: Denmark, Finland, France, Germany, Italy, Netherlands, Norway, Sweden, Spain, and the United Kingdom.

that differences in income concept definitions, i.e. individual income (Checchi et al., 2016) vs. household income (Palomino et al., 2019; Suárez Álvarez and López Menéndez, 2021), may also contribute to observed lack of consistency. However, as we will detail in the next paragraph, regardless of their income definition all of these studies are only moderately correlated with IGE rankings that are calculated with respect to individual incomes. This pattern does not support an alternative explanation based on differences in income definitions (see Panel [b] of Table 4).

TABLE 4 – Rank Correlations of Existing Studies

	Existing Studies			This Paper	
	Checchi et al. (2016)	Palomino et al. (2019)	Suarez et al. (2021)	Tree	Forest
Panel (a): Equality of Opportunity (23 countries)					
Tree	.	.	.	1.000	.
Forest	.	.	.	0.984	1.000
Checchi et al. (2016)	1.000	.	.	0.363	0.345
Palomino et al. (2019)	0.281	1.000	.	0.882	0.859
Suarez et al. (2021)	0.090	0.855	1.000	0.756	0.757
Panel (b): Intergenerational Elasticity (10 countries)					
Stuhler (2018) & Carmichael et al. (2020)	0.535	0.657	0.444	0.900	0.887

Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This table shows country rank correlations in inequality of opportunity estimates across existing studies. Panel (a) is based on the intersection of countries included in this paper, Palomino et al. (2019), Checchi et al. (2016), and Suárez Álvarez and López Menéndez (2021) (23 countries). All ranks are calculated from the mean log deviation of the counterfactual distribution $\hat{y}^C$. Panel (b) is based on the intersection of countries included in this paper, Palomino et al. (2019), Checchi et al. (2016), and Suárez Álvarez and López Menéndez (2021), and the union of Stuhler (2018) and Carmichael et al. (2020) (10 countries). Ranks in Stuhler (2018) and Carmichael et al. (2020) are calculated from consensus estimates of the intergenerational earnings elasticity (IGE). All rank correlations are based on Spearman's ρ.

In Panel (b) of Table 4 we test for conformity of opportunity rankings with the IGE literature. Inequality of opportunity rankings of existing studies are only moderately correlated with IGE rankings. In fact, various findings contradict comparative evidence on the IGE (Carmichael et al., 2020; Bratberg et al., 2017). For example, Palomino et al. (2019) and Suárez Álvarez and López Menéndez (2021) find inequality of opportunity in Germany to be at par with the Nordic countries. Checchi et al. (2016) find the Netherlands to be in the lower part of the opportunity ranking. To the contrary, rankings based on trees and forests strongly increase conformity with IGE estimates and therefore yield results that are more strongly in line with common priors about mobility in Europe.

We conclude that regression trees and forests foster consistency in the inequality of opportunity literature by reducing researcher discretion and increase conformity with evidence from the neighboring IGE literature. Both findings further bolster confidence in the ability of trees and forests to make reliable distinctions among high and low opportunity societies in Europe.

5 CONCLUSION

In this paper we propose the use of conditional inference trees and forests to estimate inequality of opportunity. Both estimation approaches minimize arbitrary model selection by the researcher while trading off downward and upward biases in inequality of opportunity estimates.

Conditional inference forests outperform all methods considered in this paper in terms of their out-of-sample prediction accuracy. This observation is valid both for simulated data generating processes and representative survey data from 31 European countries. Hence, within a given data environment they provide estimates of inequality of opportunity that have the lowest expected bias. Conditional inference trees closely mirror forests in terms of their out-of-sample prediction accuracy and their inequality of opportunity estimates. Hence, they provide a fair first-order approximation to the least biased inequality of opportunity estimates. Nevertheless, researchers should be conscious that trees may yield suboptimal results in applications with smaller samples.

We note that the improvements of our preferred methods are conditional on a given data environment. As a consequence, they do not address two major challenges of the existing literature: bias due to unobserved circumstance information, and bias due to small sample sizes. These challenges exist independent of the chosen estimation technique and can only be overcome through the availability of improved data sources in the future.

Next to their advantages, we acknowledge two potential drawbacks of our preferred methods for empirical research. First, (non-)parametric estimation approaches can be estimated by OLS—one of the workhorse estimation methods in economics and other social sciences. To the contrary, machine learning tools may require some upfront investment of applied researchers to familiarize themselves with these methods. However, as evidenced by the large volume of recent review articles, machine learning methods are increasingly integrated into the statistical toolkit of economists (Varian, 2014; Mullainathan and Spiess, 2017; Athey, 2018). Therefore, we expect this drawback to vanish over time. Second, trees and forests are computationally more costly than predictions via OLS regressions. However, in our empirical application trees approach the computation times of the (non-)parametric approach.[22] Therefore, time-constrained researchers who are willing to settle for a fair first-order approximation of the least biased method may consider using trees instead of forests.

The development of machine learning algorithms and their integration into the analytical toolkit of economists is a dynamic process. Finding the best machine learning algorithm for inequality of opportunity estimations is a methodological horse race that eventually will lead to some method outperforming the ones employed in this work. Therefore, the main contribution of this work should be understood as paving the way for new methods that are able to handle the intricacies of model selection for inequality of opportunity estimations. A particularly interesting extension may be the application of local linear forests that outperform more traditional forest algorithms in their ability to capture the linear impact of predictor variables (Friedberg et al., 2020).

Finally, we restricted ourselves to ex-ante utilitarian measures of inequality of opportunity. The exploration of these algorithms for other measurement approaches in the inequality of opportunity literature provides another interesting avenue for future research (Lefranc et al., 2009; Kanbur and Snell, 2018; Pistolesi, 2009; Brunori and Neidhöfer, 2021).

[22]The simulation of section 4.3 has the following computation times: 0.5 min (parametric), 0.5 min (non-parametric), 121.4 min (LCA), 2.1 min (trees), 2,816.2 min (forests). These computation times are based on a machine with a AMD Ryzen 7 4700U Processor (8 cores) and 16 GB RAM working memory.

References

ALESINA, Alberto, Stefanie STANTCHEVA, and Edoardo TESO (2018). "Intergenerational Mobility and Preferences for Redistribution". *American Economic Review* 108 (2), pp. 521–554.

ANDREOLI, Francesco and Alessio FUSCO (2019). "Robust cross-country analysis of inequality of opportunity". *Economics Letters* 182, pp. 86–89.

ARNESON, Richard J. (2018). "Four Conceptions of equal opportunity". *Economic Journal* 128 (612), F152–F173.

ATHEY, Susan (2018). "The Impact of Machine Learning on Economics". *The Economics of Artificial Intelligence: An Agenda*. Ed. by Ajay K. AGRAWAL, Joshua GANS, and Avi GOLDFARB. Chicago: University of Chicago Press. Chap. 21.

BIAU, Gérard and Erwan SCORNET (2016). "A random forest guided tour". *Test* 25 (2), pp. 197–227.

BJÖRKLUND, Anders, Markus JÄNTTI, and John E. ROEMER (2012). "Equality of opportunity and the distribution of long-run income in Sweden". *Social Choice and Welfare* 39 (2-3), pp. 675–696.

BLACKBURN, McKinley L. (2007). "Estimating wage differentials without logarithms". *Labour Economics* 14 (1), pp. 73–98.

BLAU, Francine D. and Lawrence M. KAHN (2017). "The Gender Wage Gap: Extent, Trends, and Explanations". *Journal of Economic Literature* 55 (3), pp. 789–865.

BLUNDELL, Jack and Erling RISA (2019). "Income and family background: Are we using the right models?" *mimeo*.

BOURGUIGNON, François, Francisco H. G FERREIRA, and Marta MENÉNDEZ (2007). "Inequality of Opportunity in Brazil". *Review of Income and Wealth* 53 (4), pp. 585–618.

BRATBERG, Espen, Jonathan DAVIS, Bhashkar MAZUMDER, Martin NYBOM, Daniel D. SCHNITZLEIN, and Kjell VAAGE (2017). "A Comparison of Intergenerational Mobility Curves in Germany, Norway, Sweden, and the US". *Scandinavian Journal of Economics* 119 (1), pp. 72–101.

BREIMAN, Leo (2001). "Random Forests". *Machine Learning* 45 (1), pp. 5–32.

BREIMAN, Leo, Jerome FRIEDMAN, Charles J. STONE, and R.A. OLSHEN (1984). *Classification and Regression Trees*. Belmont: Taylor & Francis.

BRUNORI, Paolo and Guido NEIDHÖFER (2021). "The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach". *Review of Income and Wealth*, Forthcoming.

BRUNORI, Paolo, Vito PERAGINE, and Laura SERLENGA (2019). "Upward and downward bias when measuring inequality of opportunity". *Social Choice and Welfare* 52, pp. 635–661.

BRZEZINSKI, Michal (2020). "The evolution of inequality of opportunity in Europe". *Applied Economics Letters* 27 (4), pp. 262–266.

CAPPELEN, Alexander W., Astri Drange HOLE, Erik Ø. SØRENSEN, and Bertil TUNGODDEN (2007). "The Pluralism of Fairness Ideals: An Experimental Approach". *American Economic Review* 97 (3), pp. 818–827.

CARMICHAEL, Fiona, Christian K. DARKO, Marco G. ERCOLANI, Ceren OZGEN, and W. Stanley SIEBERT (2020). "Evidence on intergenerational income transmission using complete Dutch population data". *Economics Letters* 189, p. 108996.

CARRANZA, Rafael (2020). "Upper and lower bound estimates of inequality of opportunity: A cross-national comparison for Europe". *ECINEQ Working Paper Series* 511.

CHECCHI, Daniele and Vito PERAGINE (2010). "Inequality of opportunity in Italy". *Journal of Economic Inequality* 8 (4), pp. 429–450.

CHECCHI, Daniele, Vito PERAGINE, and Laura SERLENGA (2016). "Inequality of Opportunity in Europe: Is There a Role for Institutions?" *Inequality: Causes and Consequences*. Ed. by Lorenzo CAPPELLARI, Solomon W. POLACHEK, and Konstantinos TATSIRAMOS. Vol. 43. Emerald Insight. Chap. 1, pp. 1–44.

CHETTY, Raj, Nathaniel HENDREN, Patrick KLINE, and Emmanuel SAEZ (2014a). "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States". *Quarterly Journal of Economics* 129 (4), pp. 1553–1623.

CHETTY, Raj, Nathaniel HENDREN, Patrick KLINE, Emmanuel SAEZ, and Nicholas TURNER (2014b). "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility". *American Economic Review* 104 (5), pp. 141–47.

CHETTY, Raj, Nathaniel HENDREN, Frina LIN, Jeremy MAJEROVITZ, and Benjamin SCUDERI (2016). "Childhood Environment and Gender Gaps in Adulthood". *American Economic Review* 106 (5), pp. 282–88.

CORAK, Miles (2013). "Income Inequality, Equality of Opportunity, and Intergenerational Mobility". *Journal of Economic Perspectives* 27 (3), pp. 79–102.

COWELL, Frank A. (2016). "Inequality and Poverty Measures". *Oxford Handbook of Well-Being and Public Policy*. Ed. by Matthew D. ADLER and Mark FLEURBAEY. Oxford: Oxford University Press. Chap. 4, pp. 82–125.

COWELL, Frank A. and Maria-Pia VICTORIA-FESER (1996). "Robustness Properties of Inequality Measures". *Econometrica* 64 (1), pp. 77–101.

DAHL, Gordon B. and Lance LOCHNER (2012). "The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit". *American Economic Review* 102 (5), pp. 1927–56.

FERREIRA, Francisco H. G and Jérémie GIGNOUX (2011). "The Measurement of Inequality of Opportunity: Theory and an Application to Latin America". *Review of Income and Wealth* 57 (4), pp. 622–657.

FRIEDBERG, Rina, Julie TIBSHIRANI, Susan ATHEY, and Stefan WAGER (2020). "Local Linear Forests". *Journal of Computational and Graphical Statistics* 30 (2), pp. 503–517.

FRIEDMAN, Jerome, Trevor HASTIE, and Robert TIBSHIRANI (2009). *The elements of statistical learning*. New York: Springer.

HOTHORN, Torsten, Kurt HORNIK, and Achim ZEILEIS (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework". *Journal of Computational and Graphical Statistics* 15 (3), pp. 651–674.

HUFE, Paul, Andreas PEICHL, John E. ROEMER, and Martin UNGERER (2017). "Inequality of Income Acquisition: The Role of Childhood Circumstances". *Social Choice and Welfare* 143 (3-4), pp. 499–544.

JAMES, Gareth, Daniela WITTEN, Trevor HASTIE, and Robert TIBSHIRANI (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

KANBUR, Ravi and Andy SNELL (2018). "Inequality Measures as Tests of Fairness". *Economic Journal* Forthcoming.

KREISMAN, Daniel and Marcos A. RANGEL (2015). "On the Blurring of the Color Line: Wages and Employment for Black Males of Different Skin Tones". *Review of Economics and Statistics* 97 (1), pp. 1–13.

LANZA, Stephanie T., Xianming TAN, and Bethany C. BRAY (2013). "Latent Class Analysis With Distal Outcomes: A Flexible Model-Based Approach". *Structural Equation Modelling* 20 (1), pp. 1–26.

LEFRANC, Arnaud, Nicolas PISTOLESI, and Alain TRANNOY (2009). “Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France”. *Journal of Public Economics* 93 (11–12), pp. 1189–1207.

LI DONNI, Paolo, Juan Gabriel RODRÍGUEZ, and Pedro ROSA DIAS (2015). “Empirical definition of social types in the analysis of inequality of opportunity: a latent classes approach”. *Social Choice and Welfare* 44 (3), pp. 673–701.

MILANOVIC, Branko (2015). “Global Inequality of Opportunity: How Much of Our Income is Determined by Where We Live?” *Review of Economics and Statistics* 97 (2), pp. 452–460.

MORGAN, James N. and John A. SONQUIST (1963). “Problems in the Analysis of Survey Data, and a Proposal”. *Journal of the American Statistical Association* 58 (302), pp. 415–434.

MULLAINATHAN, Sendhil and Jann SPIESS (2017). “Machine Learning: An Applied Econometric Approach”. *Journal of Economic Perspectives* 31 (2), pp. 87–106.

PALOMINO, Juan C., Gustavo A. MARRERO, and Juan G. RODRÍGUEZ (2019). “Channels of inequality of opportunity: The role of education and occupation in Europe”. *Social Indicators Research* 143, pp. 1045–1074.

PISTOLESI, Nicolas (2009). “Inequality of opportunity in the land of opportunities, 1968–2001”. *Journal of Economic Inequality* 7 (4), pp. 411–433.

RAMOS, Xavier and Dirk VAN DE GAER (2016). “Empirical Approaches to Inequality of Opportunity: Principles, Measures, and Evidence”. *Journal of Economic Surveys* 30 (5), pp. 855–883.

ROEMER, John E. (1998). *Equality of Opportunity*. Cambridge: Harvard University Press.

ROEMER, John E. and Alain TRANNOY (2015). “Equality of Opportunity”. *Handbook of Income Distribution*. Ed. by Anthony B. ATKINSON and François BOURGUIGNON. Vol. 2A. Amsterdam: Elsevier. Chap. 4, pp. 217–300.

SCHAFER, Joseph L (1999). “Multiple imputation: a primer”. *Statistical Methods in Medical Research* 8 (1), pp. 3–15.

STUHLER, Jan (2018). *A Review of Intergenerational Mobility and its Drivers*. Luxembourg: Publications Office of the European Union.

SUÁREZ ÁLVAREZ, Ana and Ana Jesús LÓPEZ MENÉNDEZ (2021). “Dynamics of inequality and opportunities within European countries”. *Bulletin of Economic Research*, Forthcoming.

VAN DE GAER, Dirk (1993). “Equality of Opportunity and Investment in Human Capital”. PhD thesis. University of Leuven.

VARIAN, Hal R. (2014). “Big Data: New Tricks for Econometrics”. *Journal of Economic Perspectives* 28 (2), pp. 3–27.

The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees and Forests

Paolo Brunori, Paul Hufe & Daniel G. Mahler

Supplementary Material

Published in *The Scandinavian Journal of Economics*

Contents

S.1 EMPIRICAL CHOICES

Tuning of Trees. Alternatively to specifying α^* a priori, it can be chosen by K-fold cross-validation (CV), which—under some minimal assumptions—provides unbiased estimates of the out-of-sample MSE (Friedman et al., 2009). First, one splits the sample into K equal-sized folds. Second, one implements the conditional inference algorithm on the union of $K-1$ folds for varying levels of α. This step makes it possible to compare the prediction from the $K-1$ folds with the unused data points in the kth fold. Third, one calculates the out-of-sample MSE as a function of α:

$$\mathrm{MSE}_k^{\mathrm{CV}}(\alpha) = \frac{1}{N^k} \sum_{i \in k} (y_i^k - \hat{f}^{-k}(\omega_i; \alpha))^2,\ \omega_i \in \Omega,\ i \in \mathcal{N}, \tag{9}$$

where $\hat{f}^{-k}()$ denotes the estimation function constructed while leaving out the kth fold. Fourth, this exercise is repeated for all K folds, so that $\mathrm{MSE}^{\mathrm{CV}}(\alpha) = \frac{1}{K} \sum_k \mathrm{MSE}_k^{\mathrm{CV}}(\alpha)$. Finally, one chooses α^* such that

$$\alpha^* = \underset{\alpha}{\operatorname{argmin}} \{\mathrm{MSE}^{\mathrm{CV}}(\alpha) : \alpha \in (0,1)\}. \tag{10}$$

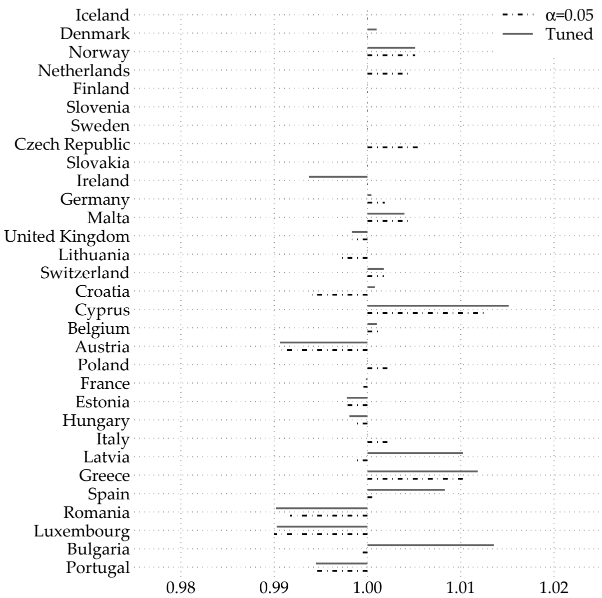
Supplementary Figure S.1 reveals that selecting α^* based on cross-validation or setting $\alpha^* = 0.05$ has little bearing on our results.

Tuning of Forests. The grid of parameters $(\alpha, \bar{P}, B)$ can be imposed a priori by the researcher or tuned to optimize the out-of-sample fit of the model. In our empirical illustration we proceed as follows. First, we fix B^* at a level at which the marginal gain of drawing an additional subsample in terms of out-of-sample prediction accuracy becomes negligible. Empirical tests show that this is the case with $B^* = 200$ for most countries in our sample (Supplementary Figure S.2).

Second, we determine α^* and $\bar{P}^*$ by minimizing the *out-of-bag* error. This entails the following three steps for a grid of values of α and $\bar{P}$:

1. Run a random forest with B^* subsamples, where $\bar{P}$ circumstances are randomly chosen to be considered at each splitting point, and α is used as the critical value for the hypothesis

FIGURE S.1 – Tuning Conditional Inference Trees



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows MSE^{Test} for different specifications of α^* relative to the baseline specification of $\alpha^* = 0.01$. Relative $\text{MSE}^{\text{Test}} > 1$ indicates worse fit than the baseline specification. Tuning is conducted by 5-fold cross-validation.

tests.

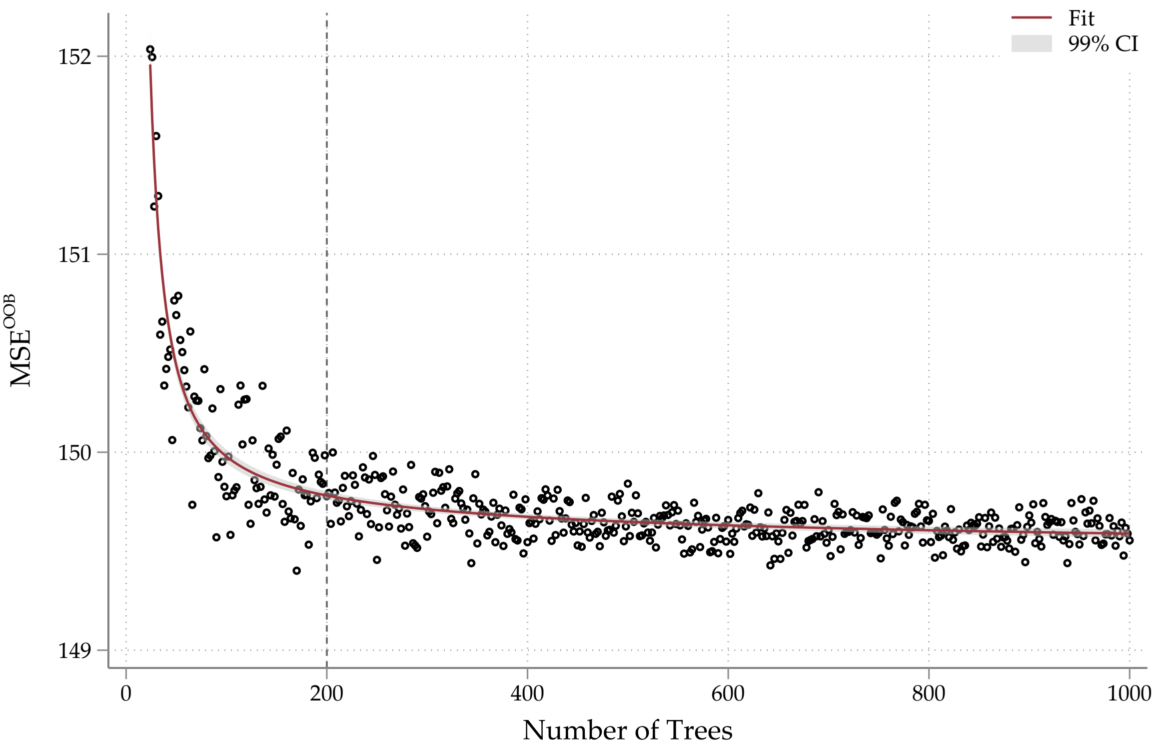
2. Calculate the average predicted value of observation i using each of the prediction functions estimated in the subsamples $\mathcal{B}_{-i} := \{\mathcal{S}' \subset \mathcal{S} : \mathcal{S}' \cap \{(y_i, \omega_i)\} = \emptyset\}$ (the so called *bags*) in which i does not enter: $\hat{f}^{\text{OOB}}(\omega_i; \alpha, \bar{P}) = \frac{1}{N_{\mathcal{B}_{-i}}} \sum_{\mathcal{S}' \in \mathcal{B}_{-i}} \hat{f}^{\mathcal{S}'}(\omega_i; \alpha, \bar{P})$.

3. Calculate the out-of-bag mean squared error:
$\text{MSE}^{\text{OOB}}(\alpha, \bar{P}) = \frac{1}{N} \sum_i [y_i - \hat{f}^{\text{OOB}}(\omega_i; \alpha, \bar{P})]^2$.

Finally, one chooses the combination of parameter values that delivers the lowest MSE^{OOB}:

$$(\alpha^*,\ \bar{P}^*) = \underset{\alpha, \bar{P}}{\operatorname{argmin}} \{\text{MSE}^{\text{OOB}} : (\alpha, \bar{P}) \in (0,1) \times \bar{\mathbf{P}}\}. \tag{11}$$

Figure S.2 – Optimal Size of Forests



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows MSE$^{\text{OOB}}$ for different specifications of B^* in the country sample of Germany. We multiply MSE$^{\text{OOB}}$ by 1×10^{-6}. We fix $\bar{P}^* = 6$. The red line is a non-parametric fit for MSE$^{\text{OOB}}$ estimates. Gray shades indicate the 99% confidence interval.

S.2 UPWARD BIAS, DOWNWARD BIAS, AND THE MSE

A standard statistic to assess prediction accuracy is the mean squared error (MSE):

$$\text{MSE} = \mathbb{E}_{\mathcal{S}}[(y - \hat{f}(\omega))^2], \tag{12}$$

where y is the observed outcome and $\hat{f}(\omega)$ the estimator of the conditional expectation $\mathbb{E}(y|\omega)$ in a random sample $\mathcal{S}$. The MSE can be decomposed into (1) variance, (2) expected bias, (3) irreducible error term (Friedman et al., 2009):

$$\text{MSE} = \text{Var}(\hat{f}(\omega)) + \mathbb{E}_{\mathcal{S}}[f(\omega) - \hat{f}(\omega)]^2 + \text{Var}(\epsilon), \tag{13}$$

$$= \underbrace{\text{Var}(f(\omega) - \hat{f}(\omega))}_{(1)} + \underbrace{(f(\omega) - \mathbb{E}_{\mathcal{S}}[\hat{f}(\omega)])^2}_{(2)} + \underbrace{\text{Var}(\epsilon)}_{(3)}. \tag{14}$$

In the literature on statistical learning this is referred to as the variance-bias decomposition. All three components can be linked to upward and downward biases in inequality of opportunity estimates.

(1) The variance captures upward bias due to model mis-specification (Brunori et al., 2019). To see this, note that we minimize (1) by imposing the following model specification $y = \hat{f}(\omega) + \epsilon = \beta_0 + \epsilon$, i.e. by assuming that individual outcomes are best predicted by the sample mean $\mu^{\mathcal{S}}$.[1] As a consequence, (1) drops out and MSE is entirely captured by components (2) and (3):

$$\begin{aligned} \text{MSE} &= \text{Var}(f(\omega) - \hat{f}(\omega)) + (f(\omega) - \mathbb{E}_{\mathcal{S}}[\hat{f}(\omega)])^2 + \text{Var}(\epsilon) \\ &= (f(\omega) - \mu)^2 + \text{Var}(\epsilon). \end{aligned}$$

Hence, *the variance-minimizing estimation model* cannot yield upward biased estimates of inequality of opportunity since it is restricted in a way that does not allow for any role of ω in the explanation of y. To the contrary, it will be downward biased. For any functional $I()$ that satisfies the measurement criteria outlined in section 2, $I(\hat{y}^C) = 0$.

[1]For the sake of exposition, we additionally assume $\mu^{\mathcal{S}} = \mu$. Obviously, this is a stark assumption. In reality, there will always be some variance in sample means as long as one does not capture the entire population.

(2) The expected bias captures downward bias due to model mis-specification. To see this, note that we minimize (2) by specifying a complex model that allows for all observable circumstances, their mutual interactions and non-linearities.[2] As a consequence, (2) drops out and MSE is entirely captured by components (1) and (3):

$$\begin{aligned}\text{MSE} &= \text{Var}(f(\omega) - \hat{f}(\omega)) + (f(\omega) - \mathbb{E}_{\mathcal{S}}[\hat{f}(\omega)])^2 + \text{Var}(\epsilon), \\ &= \text{Var}(f(\omega) - \hat{f}(\omega)) + \text{Var}(\epsilon).\end{aligned}$$

Hence, in expectation and within a given data environment *the bias-minimizing estimation model* cannot yield downward biased estimates of inequality of opportunity. To the contrary, it will be upward biased. The conditional expectations within a particular sample $\mathcal{S}$ is estimated with error and measurement error inflates the variance of $\hat{y}^C$ in comparison to the underlying truth: $\text{Var}(\hat{y}^C) = \text{Var}(y^C) + \text{Var}(u)$. For any functional $I()$ that satisfies the measurement criteria outlined in section 2, $I(\hat{y}^C) > I(y^C)$.

(3) The irreducible error term contains downward bias due to unobserved circumstance variables. To see this, assume we relax the assumption that we observe the full set of relevant circumstances. In this case, variation due to unobserved circumstances is captured in the irreducible error term (3). This part of downward bias prevails regardless of estimation method and decreases as more circumstance information becomes available. Therefore, minimizing the out-of-sample MSE corresponds to minimizing expected bias in inequality of opportunity estimates *conditional on a given data environment*.

[2] For the sake of exposition, we additionally assume that we observe all relevant circumstances. Obviously, this is a stark assumption. In reality, non-observable circumstance information is a key reason for downward bias in inequality of opportunity estimates that prevails regardless of estimation methods.

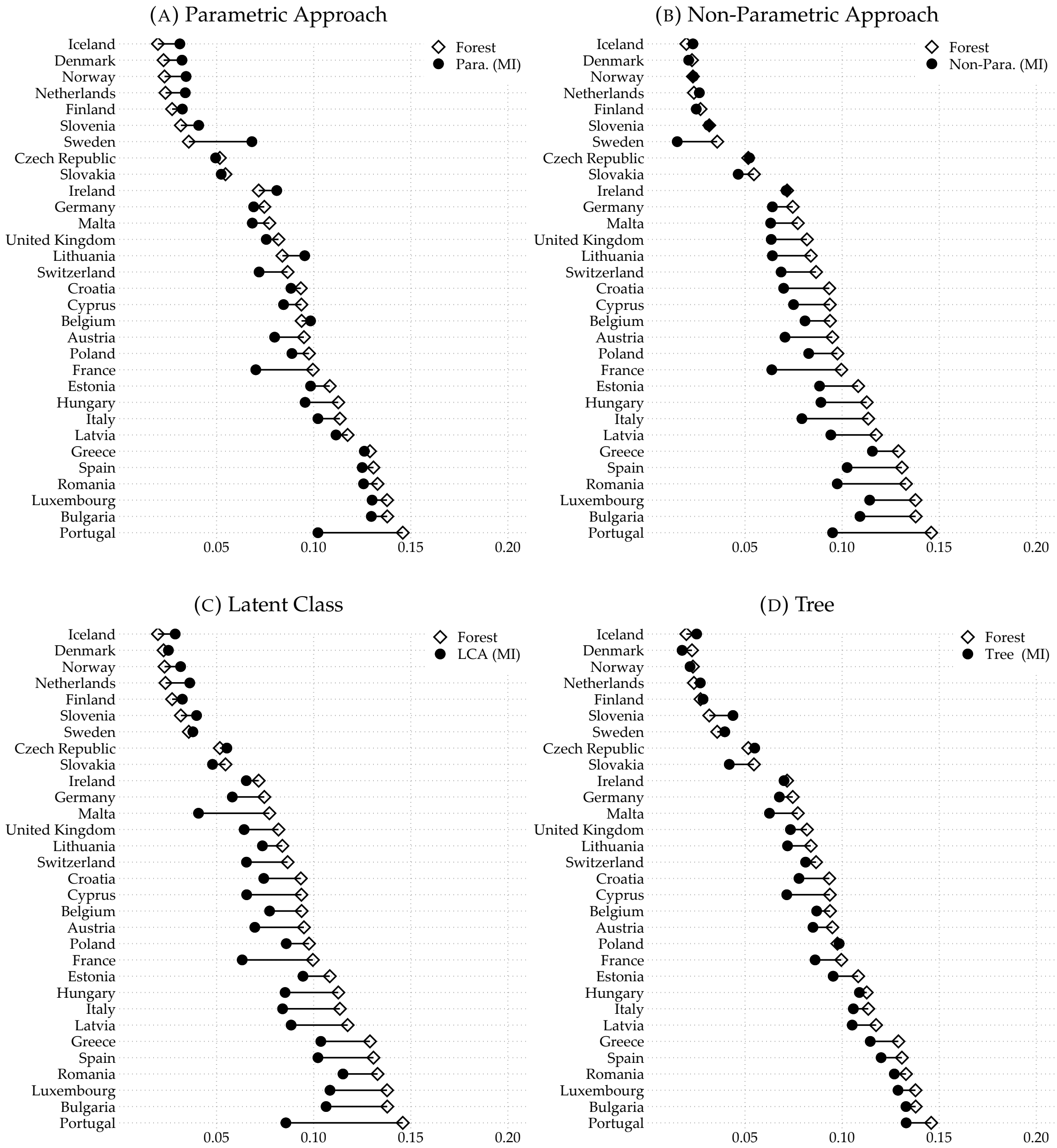
S.3 SENSITIVITY TO SAMPLE SIZE

TABLE S.1 – Sample Size by Method

Country	Benchmark Methods			Conditional Inference	
	Parametric	Non-Parametric	Latent Class	Tree	Forest
Austria	6,060	6,107	6,042	6,220	6,220
Belgium	5,289	5,439	4,528	6,011	6,011
Bulgaria	6,100	6,208	5,949	7,146	7,146
Croatia	5,997	6,159	5,945	6,945	6,945
Cyprus	4,491	4,525	4,483	4,589	4,589
Czech Republic	6,488	6,524	6,438	8,711	8,711
Denmark	2,153	2,230	2,054	5,795	5,795
Estonia	4,918	5,004	4,857	5,338	5,338
France	10,214	10,433	10,104	11,078	11,078
Germany	10,964	11,149	10,539	12,683	12,683
Greece	5,767	5,862	5,743	6,184	6,184
Hungary	12,324	12,526	12,139	13,330	13,330
Iceland	1,479	1,550	1,445	3,682	3,682
Ireland	3,102	3,164	3,080	4,318	4,318
Italy	20,284	20,803	20,238	21,070	21,070
Latvia	6,142	6,192	6,046	6,423	6,423
Lithuania	4,613	4,705	4,539	5,403	5,403
Luxembourg	6,567	6,654	6,528	6,765	6,765
Malta	4,082	4,117	4,048	4,701	4,701
Netherlands	5,461	5,598	5,414	11,411	11,411
Norway	2,355	2,456	2,329	5,026	5,026
Poland	12,808	13,369	12,676	15,545	15,545
Portugal	5,696	5,809	5,689	5,899	5,899
Romania	5,801	6,111	5,668	7,820	7,820
Slovakia	6,212	6,404	6,170	6,779	6,779
Slovenia	4,696	4,749	4,691	13,183	13,183
Spain	14,672	14,817	14,640	15,481	15,481
Sweden	531	624	467	6,599	6,599
Switzerland	6,482	6,766	6,420	7,583	7,583
United Kingdom	5,847	5,922	5,756	7,391	7,391

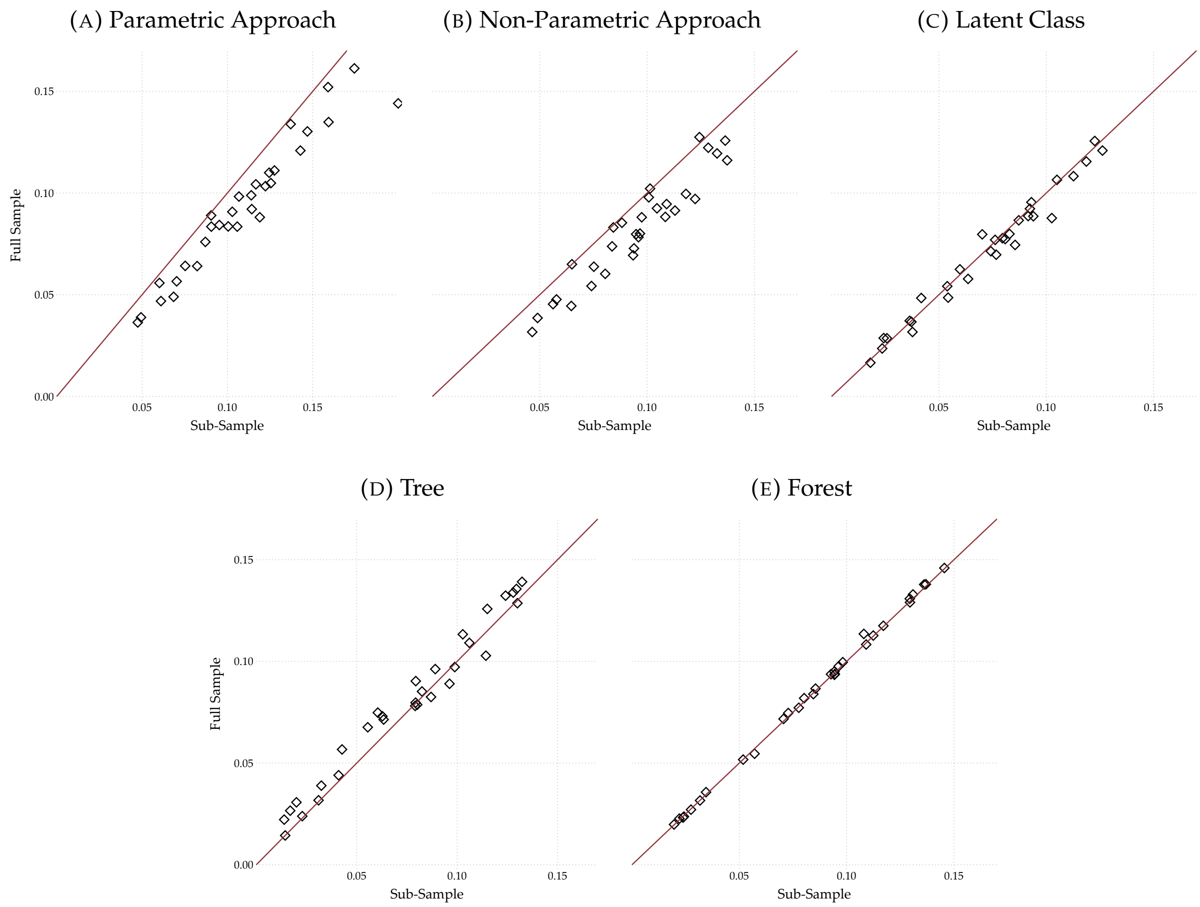
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This table shows effective sample sizes for inequality of opportunity estimations by estimation method.

FIGURE S.1 – Inequality of Opportunity Estimates: Robustness to Multiple Imputation



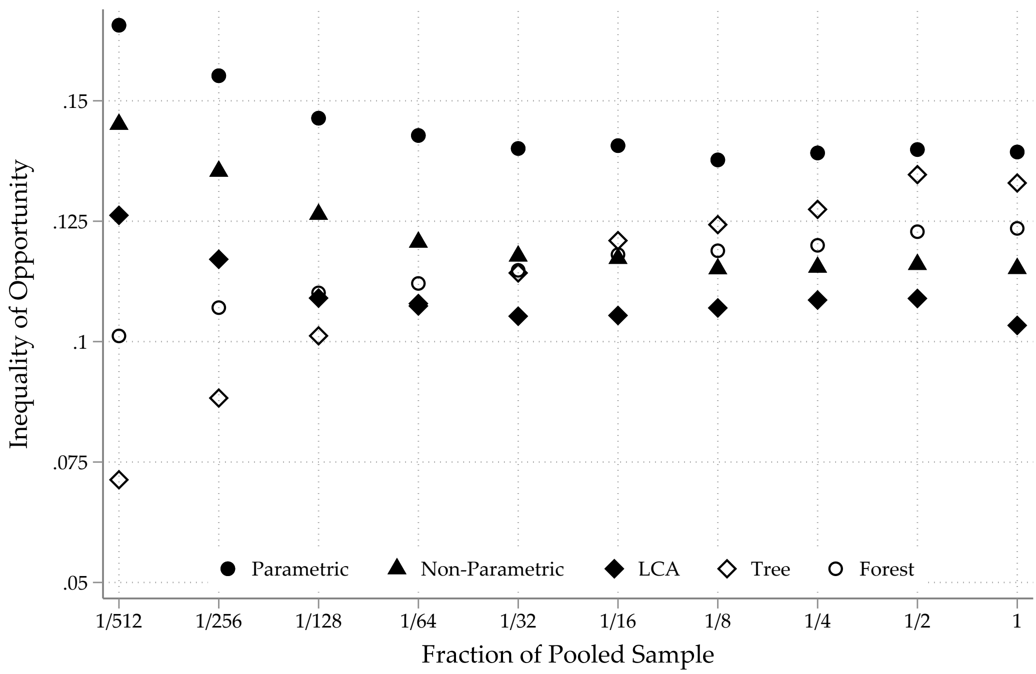
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows inequality of opportunity estimates from different estimation methods relative to forests. We impute missing circumstance information by multiple imputation such that sample sizes are constant across methods. For each country we make 10 imputations, estimate inequality of opportunity, and calculate the corresponding average. Inequality of opportunity is measured by the Gini coefficient of the counterfactual distribution $\hat{y}^C$.

FIGURE S.2 – Inequality of Opportunity Estimates: Robustness to Sample Size Reductions by Country



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows changes in inequality of opportunity estimates when reducing estimation samples to the smallest method-specific sample size. For each country-method cell we make 10 random draws from the full country sample, estimate inequality of opportunity, and calculate the corresponding average. Inequality of opportunity is measured by the Gini coefficient of the counterfactual distribution $\hat{y}^C$.

FIGURE S.3 – Inequality of Opportunity Estimates: Robustness to Sample Size Reductions (Pooled Sample)



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows changes in inequality of opportunity estimates when reducing the pooled sample to a fraction of its original size. For each fraction $1/n$ we make n random draws from the full pooled sample, estimate inequality of opportunity, and calculate the corresponding average. Inequality of opportunity is measured by the Gini coefficient of the counterfactual distribution $\hat{y}^C$.

S.4 POINT ESTIMATES AND STANDARD ERRORS

TABLE S.1 – MSE^{Test} Estimates

Country	Benchmark Methods			Conditional Inference	
	Parametric	Non-Parametric	Latent Class	Tree	Forest
Austria	188.9 [12.6]	186.5 [11.3]	184.8 [11.3]	182.5 [11.2]	178.3 [10.9]
Belgium	109.1 [7.9]	107.1 [6.7]	107.9 [6.9]	107.0 [7.3]	102.2 [7.0]
Bulgaria	5.2 [0.4]	5.2 [0.4]	5.1 [0.4]	4.9 [0.4]	4.9 [0.4]
Croatia	13.3 [0.6]	12.3 [0.5]	12.3 [0.5]	12.5 [0.6]	12.2 [0.5]
Cyprus	119.6 [8.7]	116.4 [8.2]	114.4 [8.5]	114.4 [8.0]	111.5 [7.9]
Czech Republic	16.7 [1.3]	16.7 [1.2]	16.1 [1.2]	16.2 [1.2]	16.0 [1.2]
Denmark	282.7 [32.2]	273.6 [30.4]	273.2 [30.5]	272.4 [29.5]	273.8 [29.9]
Estonia	16.6 [1.1]	15.1 [1.0]	15.5 [1.0]	15.2 [1.0]	14.7 [1.0]
Europe	253.1 [3.9]	235.3 [3.3]	235.0 [3.4]	216.3 [3.3]	210.3 [3.4]
Finland	216.5 [16.2]	211.8 [15.5]	204.8 [14.8]	206.4 [14.8]	202.3 [14.6]
France	212.3 [13.2]	206.5 [12.3]	204.4 [12.2]	203.2 [12.1]	198.9 [11.9]
Germany	149.4 [7.7]	145.2 [6.9]	145.8 [7.1]	144.7 [6.9]	142.9 [6.8]
Greece	79.3 [6.3]	68.7 [5.5]	68.3 [5.5]	67.7 [5.4]	66.8 [5.3]
Hungary	7.3 [0.3]	7.2 [0.3]	7.1 [0.3]	7.0 [0.3]	6.9 [0.3]
Iceland	92.9 [13.0]	93.9 [12.4]	89.4 [12.3]	89.9 [12.3]	91.9 [12.6]
Ireland	196.1 [16.2]	190.6 [14.6]	189.0 [14.3]	198.3 [15.0]	186.3 [14.1]
Italy	138.3 [4.4]	130.4 [3.9]	129.7 [3.9]	127.1 [3.9]	125.8 [3.9]
Latvia	13.4 [0.8]	12.3 [0.6]	12.4 [0.7]	12.3 [0.6]	11.9 [0.6]
Lithuania	9.5 [0.6]	8.9 [0.5]	8.8 [0.5]	8.8 [0.5]	8.6 [0.5]
Luxembourg	348.5 [23.4]	338.5 [21.1]	344.1 [21.1]	337.8 [20.5]	326.8 [20.1]
Malta	41.1 [3.4]	40.4 [3.1]	40.6 [3.2]	40.1 [3.1]	38.9 [3.1]
Netherlands	126.9 [7.9]	124.6 [7.2]	124.4 [7.3]	124.4 [7.2]	122.4 [7.0]
Norway	288.0 [25.2]	284.0 [22.4]	274.8 [22.4]	281.7 [22.7]	277.9 [22.4]
Poland	12.7 [0.7]	12.3 [0.6]	12.3 [0.6]	12.3 [0.6]	12.0 [0.6]
Portugal	45.2 [3.4]	44.3 [2.9]	44.2 [2.9]	44.2 [3.0]	42.0 [2.9]
Romania	2.2 [0.1]	2.2 [0.1]	2.1 [0.1]	2.1 [0.1]	2.0 [0.1]
Slovakia	12.7 [0.7]	12.4 [0.7]	12.4 [0.7]	12.4 [0.7]	12.3 [0.7]
Slovenia	34.5 [2.2]	33.6 [2.0]	33.3 [2.0]	33.4 [2.0]	33.4 [1.9]
Spain	104.1 [4.0]	97.7 [3.5]	98.5 [3.6]	97.4 [3.4]	95.2 [3.3]
Sweden	179.4 [30.6]	147.1 [27.2]	136.2 [27.6]	135.3 [28.5]	132.6 [27.7]
Switzerland	581.0 [38.7]	561.4 [35.2]	565.6 [36.3]	566.6 [35.9]	546.5 [34.8]

Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This tables shows MSE^{Test} for different estimation methods. For all methods, we multiply MSE^{Test} by 1×10^{-6}. Standard errors are derived from 200 bootstrapped re-samples of the test data.

TABLE S.2 – Inequality of Opportunity Estimates

Country	Benchmark Methods: Parametric	Benchmark Methods: Non-Parametric	Benchmark Methods: Latent Class	Conditional Inference: Tree	Conditional Inference: Forest
Austria	0.083 [0.003]	0.073 [0.004]	0.077 [0.006]	0.080 [0.003]	0.095 [0.003]
Belgium	0.098 [0.003]	0.088 [0.003]	0.077 [0.004]	0.085 [0.002]	0.094 [0.003]
Bulgaria	0.152 [0.005]	0.128 [0.005]	0.121 [0.005]	0.139 [0.002]	0.138 [0.004]
Croatia	0.110 [0.005]	0.088 [0.005]	0.080 [0.005]	0.090 [0.002]	0.093 [0.004]
Cyprus	0.091 [0.004]	0.083 [0.005]	0.071 [0.006]	0.079 [0.004]	0.094 [0.003]
Czech Republic	0.064 [0.004]	0.064 [0.004]	0.048 [0.005]	0.057 [0.001]	0.052 [0.002]
Denmark	0.047 [0.005]	0.045 [0.005]	0.017 [0.003]	0.022 [0.002]	0.023 [0.002]
Estonia	0.105 [0.006]	0.097 [0.005]	0.092 [0.007]	0.097 [0.009]	0.108 [0.005]
Finland	0.049 [0.005]	0.045 [0.004]	0.029 [0.005]	0.031 [0.008]	0.027 [0.002]
France	0.083 [0.003]	0.078 [0.004]	0.075 [0.005]	0.089 [0.004]	0.100 [0.003]
Germany	0.064 [0.003]	0.060 [0.003]	0.058 [0.003]	0.068 [0.001]	0.075 [0.002]
Greece	0.144 [0.009]	0.122 [0.006]	0.107 [0.005]	0.126 [0.006]	0.129 [0.005]
Hungary	0.104 [0.003]	0.098 [0.003]	0.089 [0.004]	0.113 [0.002]	0.113 [0.002]
Iceland	0.036 [0.007]	0.032 [0.005]	0.024 [0.005]	0.014 [0.001]	0.020 [0.002]
Ireland	0.099 [0.006]	0.093 [0.007]	0.063 [0.008]	0.078 [0.004]	0.072 [0.004]
Italy	0.111 [0.003]	0.085 [0.002]	0.089 [0.004]	0.109 [0.001]	0.114 [0.002]
Latvia	0.121 [0.007]	0.100 [0.006]	0.087 [0.006]	0.103 [0.006]	0.118 [0.004]
Lithuania	0.088 [0.008]	0.069 [0.006]	0.070 [0.007]	0.071 [0.006]	0.084 [0.005]
Luxembourg	0.134 [0.003]	0.120 [0.003]	0.116 [0.005]	0.134 [0.003]	0.138 [0.003]
Malta	0.084 [0.004]	0.080 [0.004]	0.054 [0.006]	0.073 [0.004]	0.077 [0.004]
Netherlands	0.056 [0.004]	0.054 [0.004]	0.032 [0.004]	0.027 [0.003]	0.024 [0.002]
Norway	0.039 [0.004]	0.039 [0.005]	0.029 [0.004]	0.024 [0.003]	0.023 [0.002]
Poland	0.103 [0.003]	0.095 [0.003]	0.088 [0.003]	0.096 [0.003]	0.098 [0.003]
Portugal	0.130 [0.006]	0.126 [0.005]	0.096 [0.006]	0.136 [0.006]	0.146 [0.005]
Romania	0.161 [0.006]	0.102 [0.005]	0.126 [0.007]	0.129 [0.004]	0.133 [0.004]
Slovakia	0.057 [0.004]	0.048 [0.003]	0.049 [0.004]	0.044 [0.003]	0.055 [0.003]
Slovenia	0.076 [0.004]	0.074 [0.004]	0.037 [0.003]	0.039 [0.001]	0.032 [0.002]
Spain	0.135 [0.004]	0.116 [0.003]	0.108 [0.005]	0.132 [0.002]	0.131 [0.003]
Sweden	0.089 [0.036]	0.065 [0.009]	0.037 [0.003]	0.032 [0.004]	0.036 [0.002]
Switzerland	0.084 [0.004]	0.080 [0.004]	0.080 [0.005]	0.083 [0.005]	0.087 [0.003]
United Kingdom	0.092 [0.004]	0.091 [0.005]	0.078 [0.007]	0.075 [0.002]	0.082 [0.003]

Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This tables shows inequality of opportunity estimates for different estimation methods. Inequality of opportunity is measured by the Gini coefficient of the counterfactual distribution $\hat{y}^C$. Standard errors are derived from 200 bootstrapped re-samples of the data.

TABLE S.3 – Differences in MSE$^{\text{Test}}$ Estimates Relative to Forests

Country	Parametric	Non-Parametric	Latent Class	Tree
Austria	10.574 [0.000]	8.138 [0.000]	6.437 [0.000]	4.192 [0.005]
Belgium	6.919 [0.000]	4.924 [0.000]	5.758 [0.005]	4.827 [0.000]
Bulgaria	0.247 [0.000]	0.290 [0.000]	0.185 [0.000]	0.025 [0.275]
Croatia	1.093 [0.000]	0.064 [0.240]	0.145 [0.120]	0.259 [0.005]
Cyprus	8.040 [0.000]	4.840 [0.000]	2.841 [0.355]	2.805 [0.020]
Czech Republic	0.739 [0.000]	0.648 [0.000]	0.069 [0.220]	0.226 [0.040]
Denmark	8.923 [0.030]	−0.213 [0.480]	−0.551 [0.160]	−1.393 [0.635]
Estonia	1.870 [0.000]	0.421 [0.005]	0.828 [0.000]	0.494 [0.000]
Europe	42.791 [0.000]	24.966 [0.000]	24.674 [0.000]	6.000 [0.000]
Finland	14.230 [0.000]	9.536 [0.000]	2.474 [0.180]	4.128 [0.020]
France	13.313 [0.000]	7.558 [0.000]	5.453 [0.000]	4.285 [0.000]
Germany	6.575 [0.000]	2.377 [0.000]	2.980 [0.005]	1.814 [0.030]
Greece	12.514 [0.000]	1.913 [0.000]	1.475 [0.000]	0.899 [0.000]
Hungary	0.399 [0.000]	0.329 [0.000]	0.242 [0.000]	0.181 [0.000]
Iceland	0.996 [0.240]	1.932 [0.080]	−2.508 [0.950]	−1.992 [0.875]
Ireland	9.864 [0.000]	4.338 [0.010]	2.764 [0.040]	12.043 [0.000]
Italy	12.582 [0.000]	4.665 [0.000]	3.907 [0.000]	1.347 [0.000]
Latvia	1.488 [0.000]	0.340 [0.000]	0.504 [0.000]	0.401 [0.000]
Lithuania	0.947 [0.000]	0.312 [0.000]	0.243 [0.030]	0.217 [0.000]
Luxembourg	21.715 [0.000]	11.730 [0.000]	17.363 [0.000]	11.065 [0.000]
Malta	2.298 [0.000]	1.553 [0.005]	1.717 [0.000]	1.284 [0.000]
Netherlands	4.458 [0.000]	2.141 [0.010]	1.977 [0.030]	1.935 [0.025]
Norway	10.159 [0.020]	6.133 [0.105]	−3.122 [0.890]	3.800 [0.175]
Poland	0.636 [0.000]	0.297 [0.000]	0.254 [0.000]	0.290 [0.000]
Portugal	3.260 [0.000]	2.276 [0.000]	2.170 [0.010]	2.258 [0.000]
Romania	0.179 [0.000]	0.151 [0.000]	0.085 [0.020]	0.074 [0.000]
Slovakia	0.316 [0.000]	0.111 [0.055]	0.046 [0.230]	0.053 [0.145]
Slovenia	1.101 [0.015]	0.267 [0.185]	−0.060 [0.450]	0.059 [0.390]
Spain	8.951 [0.000]	2.560 [0.000]	3.289 [0.000]	2.215 [0.000]
Sweden	46.768 [0.000]	14.554 [0.010]	3.578 [0.230]	2.728 [0.315]
Switzerland	34.516 [0.000]	14.854 [0.005]	19.105 [0.000]	20.063 [0.000]

Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This tables shows MSE$^{\text{Test}}$ for different estimation methods relative to conditional inference forests. For all methods, we multiply MSE$^{\text{Test}}$ by 1×10^{-6} and subtract MSE$^{\text{Test}}$ of forests. *p*-values (in brackets) for one-sided tests whether MSE$^{\text{Test}}$ is smaller than MSE$^{\text{Test}}$ of forests are derived from 200 bootstrapped re-samples of the test data.

S.5 DESCRIPTIVE STATISTICS

TABLE S.4 – Descriptive Statistics (Individual and Household)

		Birth Area		Parents in HH		HH Composition			
Country	Male	Native	EU	Both	None	Adults	Working Ad.	Children	Home Owner
Austria	0.504	0.791	0.070	0.856	0.017	2.729	1.759	2.601	0.585
Belgium	0.498	0.824	0.076	0.855	0.019	2.377	1.594	2.764	0.750
Bulgaria	0.500	0.994	0.001	0.904	0.012	2.436	2.014	2.050	0.910
Croatia	0.501	0.875	0.017	0.874	0.020	2.564	1.352	2.308	0.902
Cyprus	0.525	0.787	0.096	0.900	0.015	2.642	1.669	2.695	0.784
Czech Republic	0.505	0.963	0.027	0.851	0.013	2.085	1.925	2.235	0.597
Denmark	0.499	0.917	0.028	0.816	0.025	2.226	2.323	2.239	0.745
Estonia	0.525	0.847	0.000	0.756	0.011	2.105	1.801	2.086	0.859
Finland	0.496	0.951	0.018	0.822	0.016	2.364	1.741	2.254	0.769
France	0.509	0.885	0.036	0.820	0.022	2.468	1.655	1.747	0.630
Germany	0.518	0.936	0.000	0.835	0.020	2.226	1.675	2.254	0.510
Greece	0.498	0.890	0.025	0.931	0.019	2.308	1.562	2.327	0.834
Hungary	0.512	0.989	0.008	0.843	0.042	2.140	1.747	2.259	0.828
Iceland	0.501	0.906	0.052	0.889	0.014	2.406	1.902	2.573	0.885
Ireland	0.524	0.783	0.149	0.893	0.078	3.169	3.203	3.174	0.727
Italy	0.504	0.881	0.041	0.900	0.011	2.588	1.630	2.413	0.682
Latvia	0.525	0.864	0.000	0.764	0.011	1.971	1.761	2.268	0.450
Lithuania	0.521	0.939	0.004	0.846	0.016	2.325	2.021	2.457	0.698
Luxembourg	0.499	0.480	0.401	0.868	0.020	2.530	1.640	2.696	0.734
Malta	0.497	0.945	0.000	0.927	0.021	3.024	1.842	2.665	0.573
Netherlands	0.499	0.863	0.023	0.871	0.020	2.100	1.531	3.186	0.577
Norway	0.489	0.909	0.041	0.906	0.014	2.019	1.769	1.829	0.921
Poland	0.505	0.999	0.000	0.890	0.015	2.706	1.966	2.439	0.643
Portugal	0.506	0.906	0.022	0.854	0.017	2.682	2.233	2.676	0.544
Romania	0.505	0.999	0.000	0.919	0.009	2.765	1.900	2.232	0.858
Slovakia	0.519	0.987	0.010	0.920	0.010	2.517	2.078	2.340	0.694
Slovenia	0.488	0.867	0.000	0.844	0.020	2.495	1.755	2.153	0.740
Spain	0.495	0.835	0.051	0.893	0.012	2.880	2.111	2.434	0.819
Sweden	0.491	0.800	0.062	0.802	0.038	2.057	1.772	2.377	0.742
Switzerland	0.504	0.677	0.181	0.838	0.017	2.522	1.885	2.511	0.544
United Kingdom	0.507	0.848	0.042	0.825	0.024	2.342	2.240	2.406	0.649

Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: Omitted circumstance expressions listed in order of the circumstance categories: "Female"; "Non-EU"; "Only Mother/Only Father/Collective House"; "House Not Owned". See also Table 1.

TABLE S.5 – Descriptive Statistics (Fathers)

	Birth Area		Citizenship		Education			Activity				Occupation (ISCO-08 1-Digit)									
Country	Native	EU	Resid.	EU	Low	Med.	High	Empl.	Self-Empl.	Un-empl.	House Work	1	2	3	4	5	6	7	8	9	Superv.
Austria	0.745	0.093	0.779	0.067	0.400	0.419	0.134	0.712	0.216	0.003	0.001	0.043	0.045	0.063	0.051	0.137	0.147	0.284	0.062	0.085	0.337
Belgium	0.748	0.100	0.762	0.093	0.491	0.199	0.178	0.699	0.179	0.007	0.002	0.068	0.126	0.104	0.084	0.054	0.057	0.209	0.127	0.041	0.278
Bulgaria	0.933	0.004	0.936	0.001	0.466	0.333	0.081	0.899	0.028	0.005	0.000	0.022	0.065	0.047	0.029	0.035	0.135	0.216	0.207	0.142	0.093
Croatia	0.822	0.006	0.834	0.004	0.464	0.312	0.063	0.763	0.103	0.037	0.016	0.025	0.041	0.088	0.036	0.072	0.049	0.214	0.103	0.228	0.129
Cyprus	0.803	0.082	0.808	0.094	0.667	0.178	0.091	0.566	0.381	0.004	0.000	0.011	0.071	0.074	0.029	0.104	0.161	0.245	0.122	0.125	0.229
Czech Republic	0.878	0.065	0.910	0.036	0.602	0.216	0.090	0.891	0.017	0.001	0.000	0.033	0.070	0.125	0.036	0.035	0.039	0.305	0.195	0.053	0.233
Denmark	0.930	0.026	0.967	0.023	0.350	0.421	0.229	0.712	0.267	0.005	0.001	0.111	0.132	0.072	0.042	0.110	0.154	0.282	0.067	0.010	0.455
Estonia	0.603	0.270	0.637	0.233	0.300	0.338	0.165	0.823	0.006	0.003	0.003	0.076	0.092	0.053	0.014	0.013	0.034	0.221	0.253	0.053	0.153
Finland	0.822	0.007	0.823	0.007	0.480	0.184	0.165	0.592	0.203	0.017	0.001	0.043	0.088	0.083	0.016	0.047	0.126	0.151	0.134	0.045	
France	0.789	0.078	0.857	0.057	0.695	0.073	0.095	0.753	0.170	0.003	0.001	0.084	0.068	0.111	0.072	0.038	0.103	0.155	0.055	0.223	0.335
Germany	0.870	0.130	0.931	0.069	0.100	0.508	0.242	0.820	0.125	0.007	0.001	0.047	0.117	0.168	0.057	0.061	0.057	0.258	0.142	0.035	0.316
Greece	0.887	0.016	0.911	0.015	0.587	0.135	0.084	0.449	0.517	0.002	0.000	0.073	0.047	0.026	0.087	0.046	0.308	0.210	0.099	0.055	0.182
Hungary	0.961	0.018	0.968	0.012	0.596	0.242	0.087	0.892	0.043	0.001	0.001	0.037	0.060	0.052	0.017	0.053	0.095	0.280	0.192	0.137	0.118
Iceland	0.905	0.060	0.907	0.058	0.338	0.471	0.142	0.648	0.321	0.002	0.000	0.121	0.128	0.072	0.023	0.094	0.172	0.219	0.093	0.046	0.570
Ireland	0.792	0.107	0.758	0.094	0.574	0.258	0.112	0.659	0.221	0.049	0.002	0.104	0.092	0.042	0.022	0.072	0.155	0.149	0.065	0.158	0.344
Italy	0.823	0.023	0.827	0.021	0.704	0.138	0.039	0.615	0.244	0.015	0.004	0.056	0.041	0.073	0.057	0.068	0.097	0.227	0.105	0.118	0.199
Latvia	0.571	0.250	0.642	0.166	0.385	0.295	0.099	0.768	0.005	0.002	0.003	0.037	0.083	0.037	0.010	0.019	0.070	0.200	0.216	0.082	0.071
Lithuania	0.899	0.004	0.926	0.004	0.538	0.228	0.085	0.916	0.011	0.000	0.001	0.049	0.074	0.038	0.017	0.023	0.080	0.241	0.179	0.214	0.110
Luxembourg	0.387	0.467	0.400	0.466	0.484	0.316	0.120	0.757	0.174	0.001	0.001	0.063	0.093	0.118	0.048	0.035	0.112	0.228	0.183	0.039	0.251
Malta	0.951	0.041	0.952	0.040	0.556	0.177	0.060	0.711	0.215	0.013	0.001	0.059	0.046	0.104	0.044	0.147	0.050	0.237	0.099	0.108	0.222
Netherlands	0.817	0.029	0.881	0.022	0.355	0.288	0.208	0.732	0.164	0.007	0.005	0.085	0.135	0.162	0.050	0.068	0.080	0.196	0.075	0.027	0.309
Norway	0.896	0.047	0.908	0.041	0.318	0.391	0.287	0.708	0.256	0.003	0.001	0.116	0.115	0.164	0.029	0.058	0.114	0.223	0.095	0.030	0.286
Poland	0.954	0.012	0.980	0.003	0.467	0.444	0.069	0.698	0.240	0.002	0.001	0.036	0.044	0.053	0.025	0.041	0.240	0.252	0.155	0.078	0.112
Portugal	0.932	0.006	0.945	0.006	0.700	0.031	0.030	0.650	0.248	0.002	0.001	0.047	0.032	0.060	0.038	0.082	0.185	0.264	0.114	0.077	0.190
Romania	0.938	0.002	0.940	0.001	0.719	0.090	0.029	0.643	0.233	0.004	0.015	0.005	0.037	0.036	0.016	0.017	0.250	0.245	0.125	0.105	0.047
Slovakia	0.935	0.020	0.945	0.011	0.362	0.497	0.075	0.921	0.011	0.002	0.001	0.042	0.060	0.095	0.028	0.043	0.030	0.285	0.209	0.128	0.145
Slovenia	0.761	0.206			0.665	0.175	0.096	0.771	0.095	0.012	0.012	0.026	0.058	0.104	0.036	0.052	0.081	0.252	0.083	0.166	0.251
Spain	0.836	0.047	0.846	0.046	0.762	0.064	0.081	0.702	0.219	0.006	0.001	0.056	0.045	0.076	0.055	0.087	0.145	0.191	0.113	0.137	0.191
Sweden	0.936	0.023	0.818	0.069	0.412	0.354	0.182	0.742	0.210	0.003	0.002	0.046	0.116	0.062	0.025	0.096	0.083	0.217	0.113	0.017	0.333
Switzerland	0.609	0.273	0.621	0.265	0.225	0.493	0.155	0.653	0.291	0.001	0.000	0.089	0.133	0.143	0.057	0.059	0.109	0.223	0.078	0.049	0.401
United Kingdom	0.800	0.064	0.869	0.039	0.508	0.228	0.150	0.795	0.147	0.025	0.002	0.095	0.142	0.085	0.040	0.075	0.036	0.236	0.133	0.083	0.398

Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: Omitted circumstance expressions listed in order of the circumstance categories: “Non-EU”; “Not Europe”; “Dead/Unknown/Illiterate”; “Dead/Unknown/Retired/Other Inactive”; "Dead/Unknown/Not Working/Armed Forces"; “Dead/Unknown/Not Working/Non-Supervisory”. Information on parental citizenship (supervisory status) are not available for Slovenia (Finland). See also Table 1.

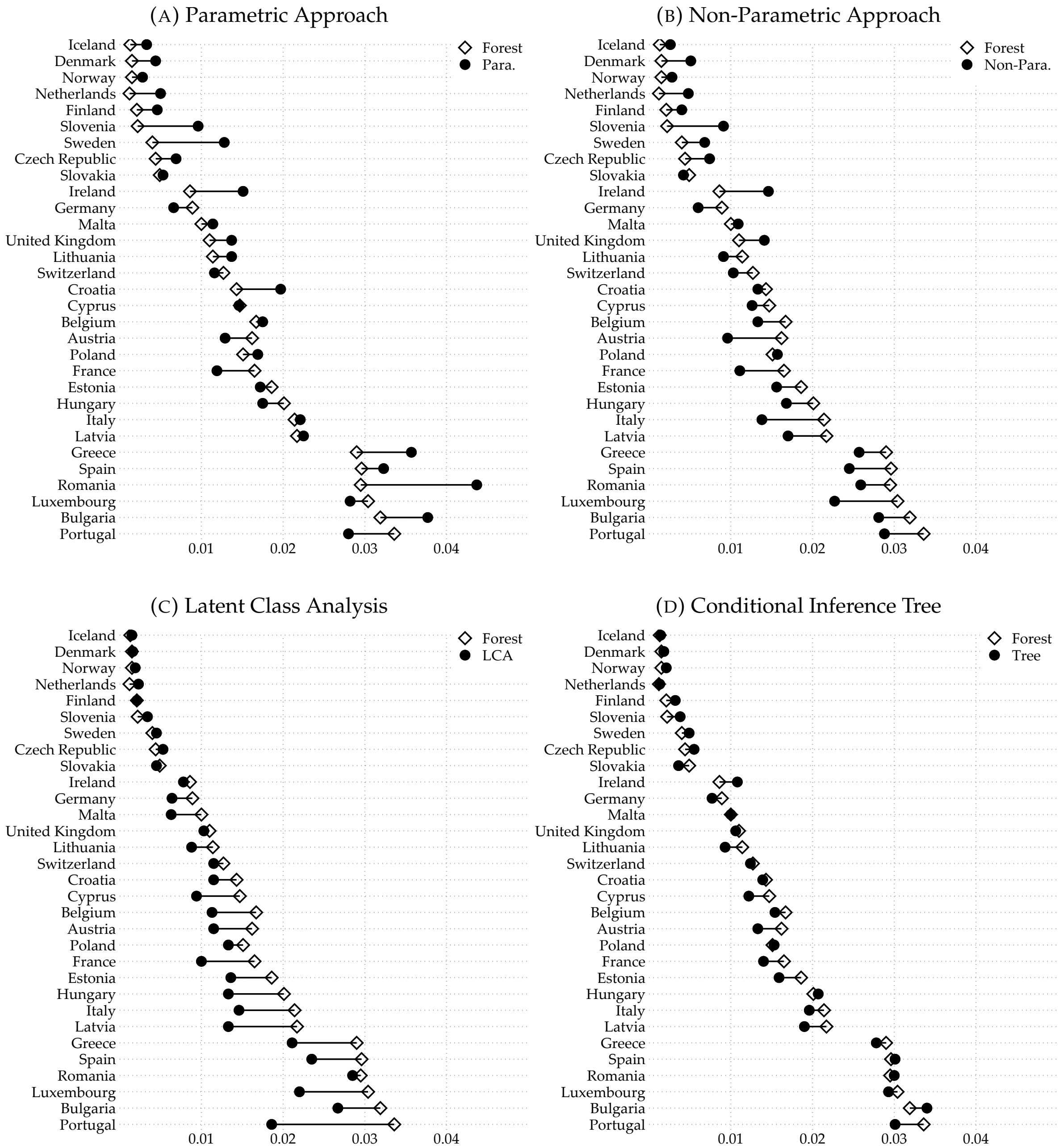
TABLE S.6 – Descriptive Statistics (Mothers)

	Birth Area		Citizenship		Education			Activity				Occupation (ISCO-08 1-Digit)									
Country	Native	EU	Resid.	EU	Low	Med.	High	Empl.	Self-Empl.	Un-empl.	House Work	1	2	3	4	5	6	7	8	9	Superv.
Austria	0.741	0.096	0.791	0.064	0.593	0.323	0.040	0.366	0.170	0.002	0.436	0.009	0.023	0.008	0.070	0.153	0.130	0.045	0.010	0.088	0.092
Belgium	0.755	0.097	0.790	0.092	0.564	0.201	0.126	0.320	0.117	0.006	0.508	0.009	0.081	0.045	0.058	0.046	0.002	0.016	0.024	0.069	0.034
Bulgaria	0.931	0.003	0.981	0.002	0.464	0.356	0.099	0.878	0.026	0.007	0.058	0.008	0.123	0.040	0.092	0.140	0.181	0.099	0.064	0.151	0.030
Croatia	0.823	0.008	0.848	0.003	0.634	0.189	0.043	0.352	0.053	0.027	0.538	0.003	0.058	0.036	0.046	0.070	0.022	0.034	0.013	0.122	0.033
Cyprus	0.804	0.080	0.812	0.091	0.684	0.162	0.057	0.325	0.166	0.001	0.492	0.001	0.046	0.020	0.037	0.067	0.036	0.022	0.042	0.220	0.048
Czech Republic	0.882	0.061	0.946	0.037	0.670	0.261	0.043	0.898	0.007	0.003	0.064	0.011	0.080	0.104	0.149	0.160	0.074	0.105	0.080	0.139	0.088
Denmark	0.916	0.032	0.929	0.026	0.517	0.289	0.194	0.641	0.068	0.006	0.266	0.020	0.111	0.096	0.126	0.228	0.031	0.050	0.026	0.001	0.123
Estonia	0.601	0.272	0.726	0.250	0.334	0.391	0.208	0.906	0.004	0.001	0.051	0.049	0.169	0.109	0.097	0.110	0.084	0.051	0.124	0.113	0.085
Finland	0.821	0.007	0.932	0.006	0.548	0.242	0.147	0.666	0.189	0.021	0.057	0.012	0.125	0.096	0.118	0.145	0.047	0.045	0.054	0.199	
France	0.806	0.067	0.880	0.047	0.724	0.079	0.072	0.454	0.085	0.001	0.424	0.011	0.036	0.050	0.111	0.107	0.059	0.049	0.005	0.109	0.072
Germany	0.878	0.122	0.937	0.063	0.262	0.507	0.090	0.486	0.052	0.008	0.431	0.013	0.055	0.086	0.097	0.118	0.023	0.015	0.078	0.031	0.062
Greece	0.888	0.016	0.916	0.016	0.592	0.133	0.044	0.193	0.277	0.001	0.517	0.023	0.027	0.004	0.039	0.048	0.223	0.034	0.021	0.049	0.026
Hungary	0.964	0.017	0.980	0.012	0.652	0.247	0.052	0.730	0.022	0.001	0.216	0.014	0.050	0.064	0.114	0.120	0.062	0.076	0.086	0.165	0.043
Iceland	0.895	0.069	0.910	0.058	0.614	0.276	0.079	0.610	0.100	0.000	0.268	0.030	0.100	0.050	0.111	0.183	0.061	0.027	0.013	0.130	0.152
Ireland	0.787	0.114	0.761	0.103	0.546	0.324	0.097	0.253	0.048	0.007	0.673	0.022	0.061	0.007	0.052	0.059	0.017	0.014	0.007	0.060	0.082
Italy	0.819	0.024	0.863	0.025	0.775	0.115	0.024	0.226	0.081	0.005	0.634	0.012	0.040	0.022	0.029	0.051	0.033	0.032	0.023	0.063	0.041
Latvia	0.584	0.236	0.792	0.182	0.419	0.397	0.123	0.891	0.003	0.002	0.064	0.033	0.136	0.084	0.098	0.120	0.085	0.092	0.024	0.220	0.075
Lithuania	0.902	0.002	0.959	0.003	0.519	0.316	0.106	0.867	0.014	0.001	0.095	0.035	0.129	0.046	0.049	0.109	0.067	0.112	0.034	0.293	0.068
Luxembourg	0.374	0.483	0.393	0.485	0.587	0.245	0.071	0.318	0.106	0.000	0.538	0.028	0.049	0.046	0.036	0.061	0.054	0.015	0.024	0.108	0.047
Malta	0.952	0.040	0.958	0.037	0.648	0.143	0.027	0.075	0.015	0.001	0.889	0.003	0.019	0.008	0.009	0.018	0.002	0.004	0.009	0.012	0.012
Netherlands	0.814	0.030	0.901	0.023	0.501	0.303	0.094	0.298	0.052	0.003	0.614	0.011	0.053	0.042	0.056	0.090	0.015	0.011	0.008	0.061	0.040
Norway	0.877	0.048	0.888	0.044	0.362	0.437	0.187	0.626	0.107	0.007	0.232	0.031	0.044	0.144	0.114	0.206	0.053	0.016	0.027	0.093	0.070
Poland	0.957	0.011	0.990	0.004	0.531	0.404	0.056	0.516	0.264	0.008	0.174	0.018	0.057	0.052	0.070	0.095	0.265	0.080	0.018	0.118	0.050
Portugal	0.928	0.008	0.950	0.007	0.631	0.029	0.028	0.359	0.197	0.003	0.383	0.016	0.031	0.017	0.025	0.074	0.158	0.059	0.032	0.145	0.048
Romania	0.937	0.001	0.941	0.001	0.724	0.114	0.019	0.366	0.218	0.006	0.306	0.001	0.033	0.025	0.026	0.049	0.217	0.075	0.039	0.078	0.011
Slovakia	0.932	0.023	0.980	0.010	0.451	0.482	0.039	0.846	0.006	0.004	0.111	0.010	0.075	0.110	0.107	0.161	0.034	0.096	0.052	0.203	0.048
Slovenia	0.785	0.182			0.729	0.157	0.068	0.595	0.064	0.004	0.283	0.007	0.055	0.104	0.087	0.089	0.055	0.064	0.006	0.192	0.094
Spain	0.836	0.046	0.849	0.046	0.802	0.048	0.040	0.186	0.069	0.001	0.719	0.010	0.025	0.010	0.021	0.059	0.028	0.021	0.008	0.071	0.029
Sweden	0.932	0.026	0.822	0.068	0.415	0.371	0.191	0.715	0.060	0.001	0.202	0.006	0.082	0.030	0.049	0.160	0.014	0.008	0.024	0.033	0.101
Switzerland	0.585	0.296	0.613	0.273	0.406	0.407	0.059	0.383	0.156	0.001	0.425	0.027	0.059	0.069	0.070	0.129	0.055	0.040	0.024	0.065	0.066
United Kingdom	0.808	0.064	0.877	0.036	0.679	0.099	0.124	0.577	0.051	0.087	0.271	0.026	0.097	0.068	0.078	0.152	0.005	0.028	0.044	0.127	0.104

Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: Omitted circumstance expressions listed in order of the circumstance categories are: “Non-EU”; “Not Europe”; “Dead/Unknown/Illiterate”; “Dead/Unknown/Retired/Other Inactive”; "Dead/Unknown/Not Working/Armed Forces"; “Dead/Unknown/Not Working/Non-Supervisory”. Information on parental citizenship (supervisory status) are not available for Slovenia (Finland). See also Table 1.

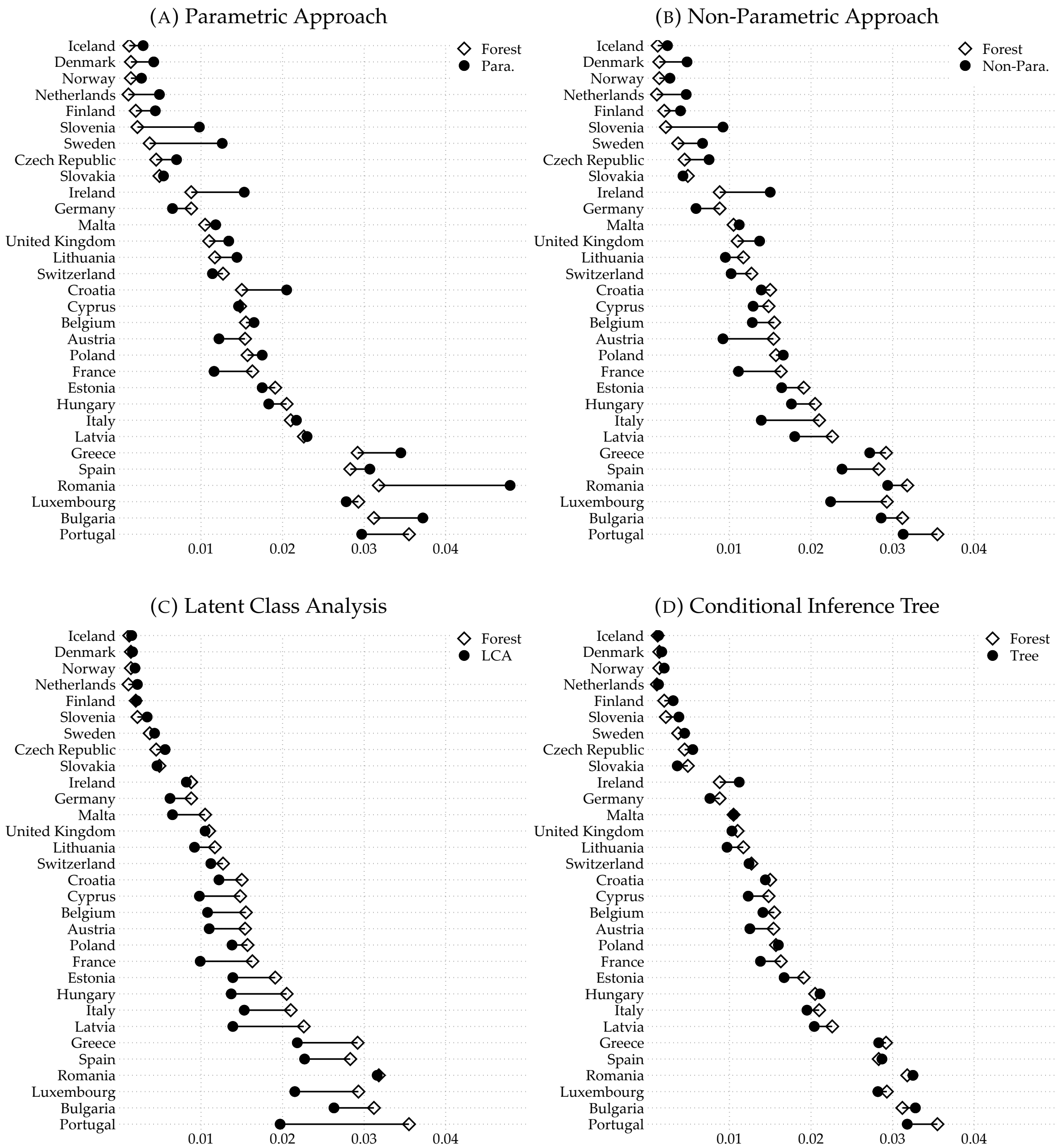
S.6 ALTERNATIVE INEQUALITY INDEXES

FIGURE S.1 – Comparison of Inequality of Opportunity Estimates by Method (GE[0])



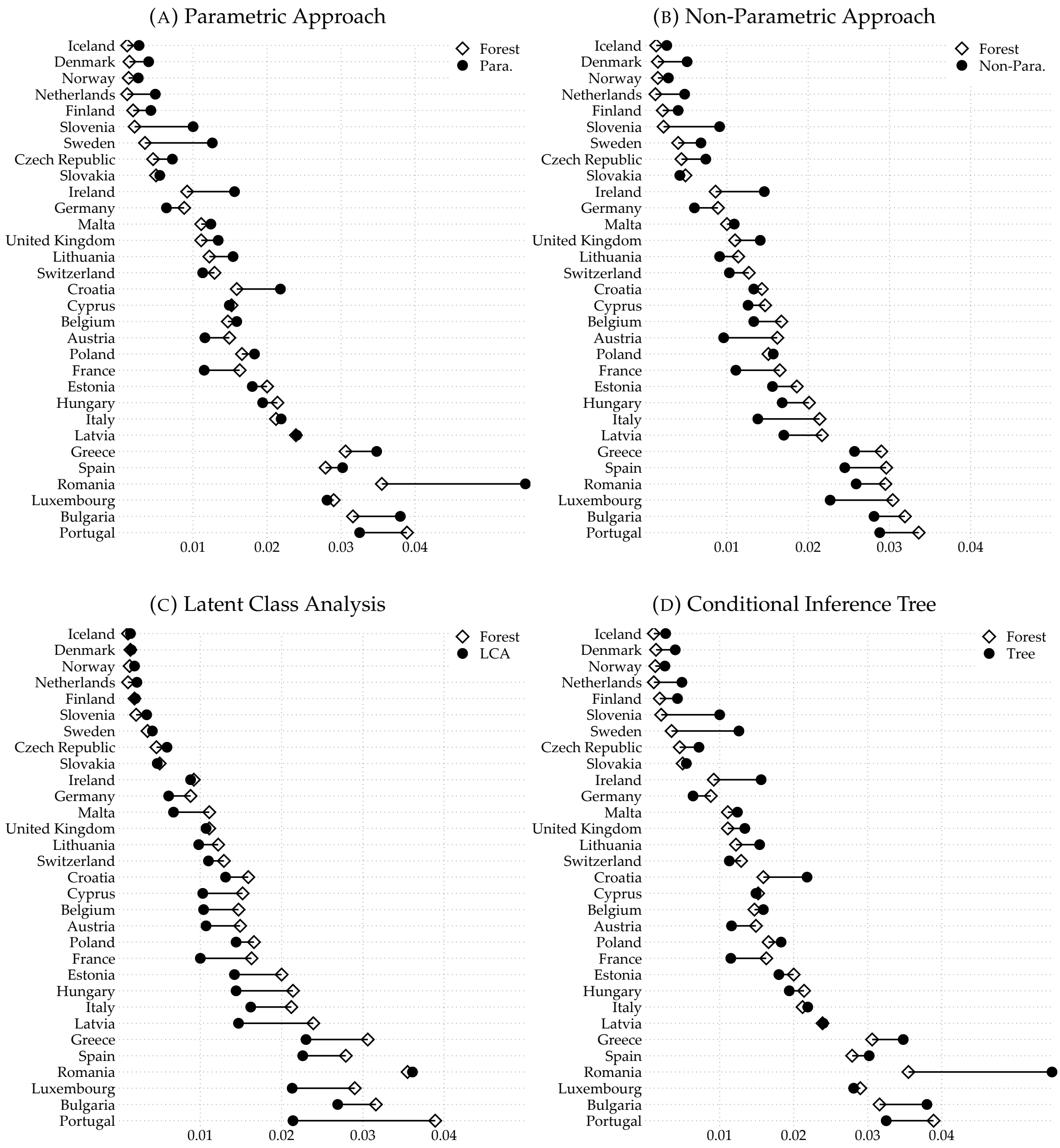
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows inequality of opportunity estimates from different estimation methods relative to random forests. Inequality of opportunity is measured by the general entropy measure ($\alpha = 0$) of the counterfactual distribution $\hat{y}^C$.

FIGURE S.2 – Comparison of Inequality of Opportunity Estimates by Method (GE[1])



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows inequality of opportunity estimates from different estimation methods relative to random forests. Inequality of opportunity is measured by the general entropy measure ($\alpha = 1$) of the counterfactual distribution $\hat{y}^C$.

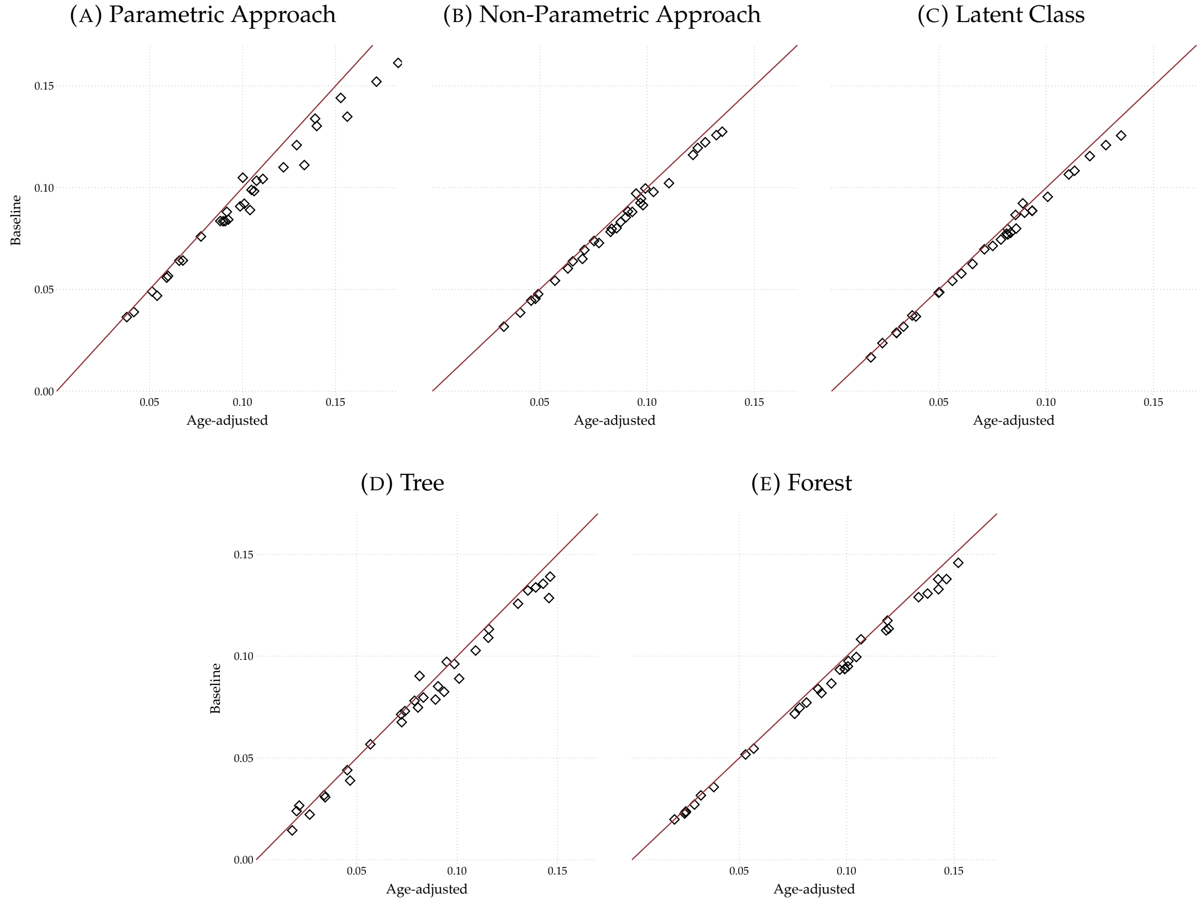
FIGURE S.3 – Comparison of Inequality of Opportunity Estimates by Method (GE[2])



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows inequality of opportunity estimates from different estimation methods relative to random forests. Inequality of opportunity is measured by the general entropy measure ($\alpha = 2$) of the counterfactual distribution $\hat{y}^C$.

S.7 ACCOUNTING FOR AGE

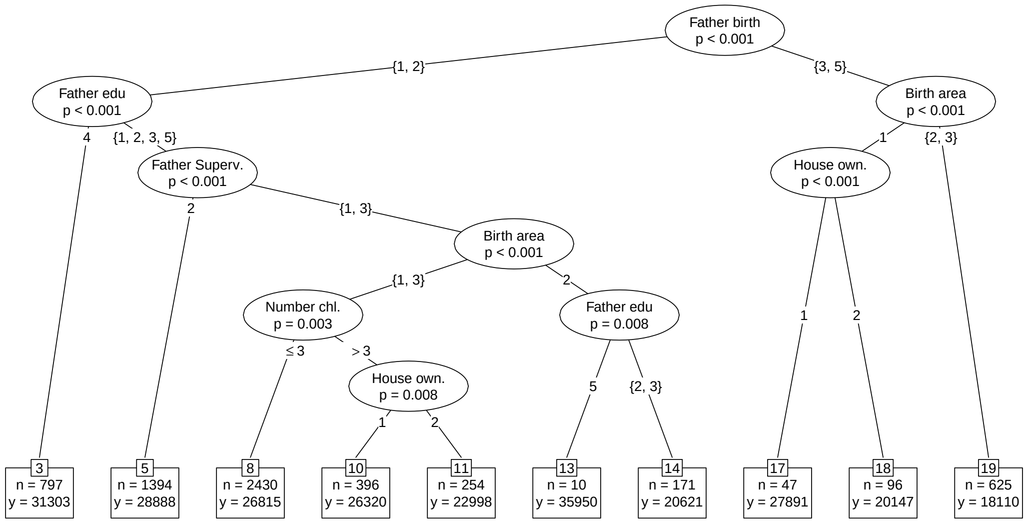
FIGURE S.4 – Inequality of Opportunity Estimates: Robustness to Age-adjusted Income Distributions



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: This figure shows changes in inequality of opportunity estimates when adjusting income distributions for age effects. For each country, we construct age-adjusted income distributions by residualizing income from age and age squared, and adding back the country-specific mean income. Inequality of opportunity is measured by the Gini coefficient of the counterfactual distribution $\hat{y}^C$.

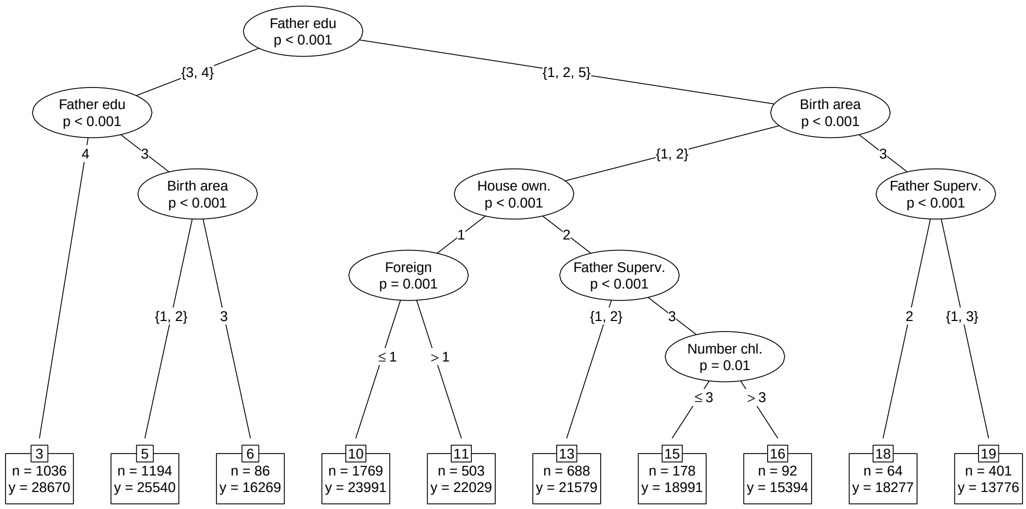
S.8 OPPORTUNITY STRUCTURES

FIGURE S.5 – Opportunity Tree (Austria)



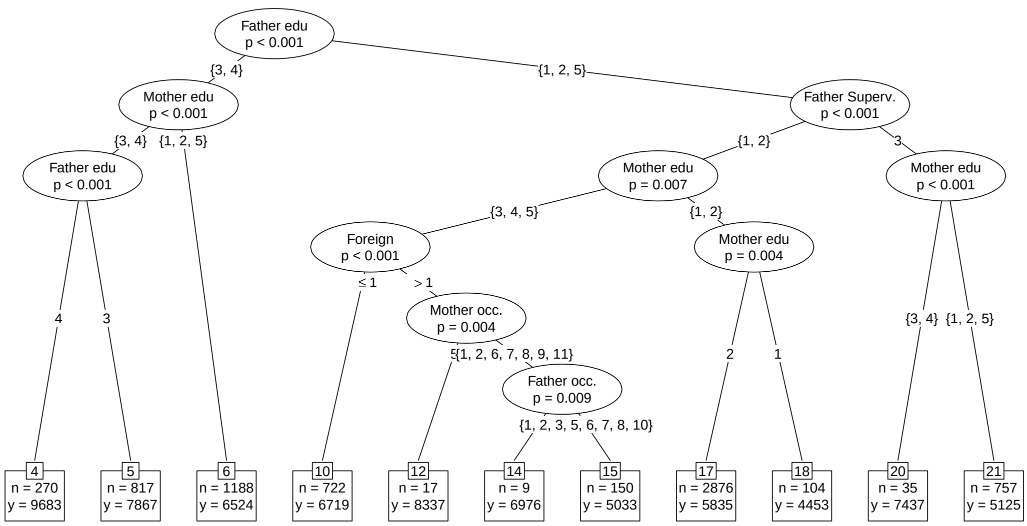
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.6 – Opportunity Tree (Belgium)



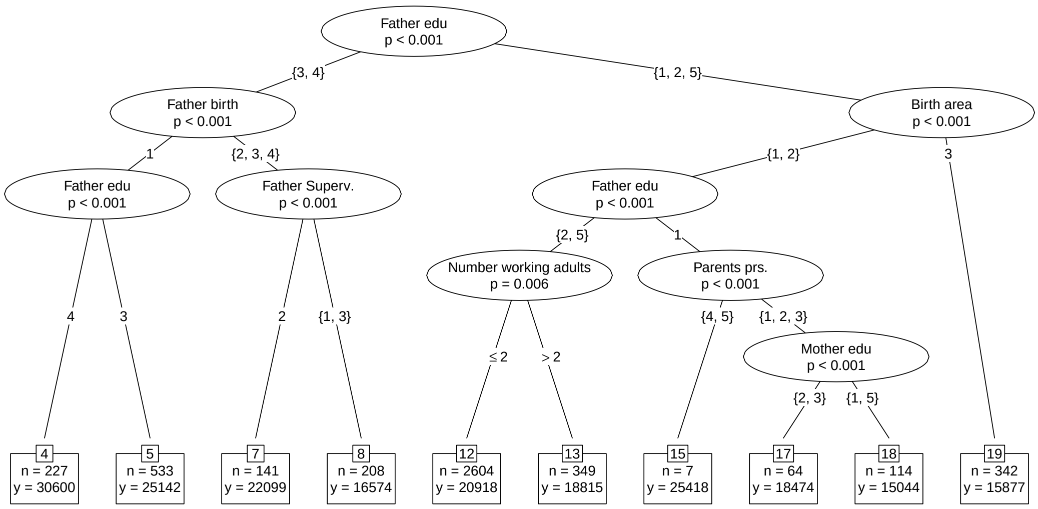
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.7 – Opportunity Tree (Croatia)



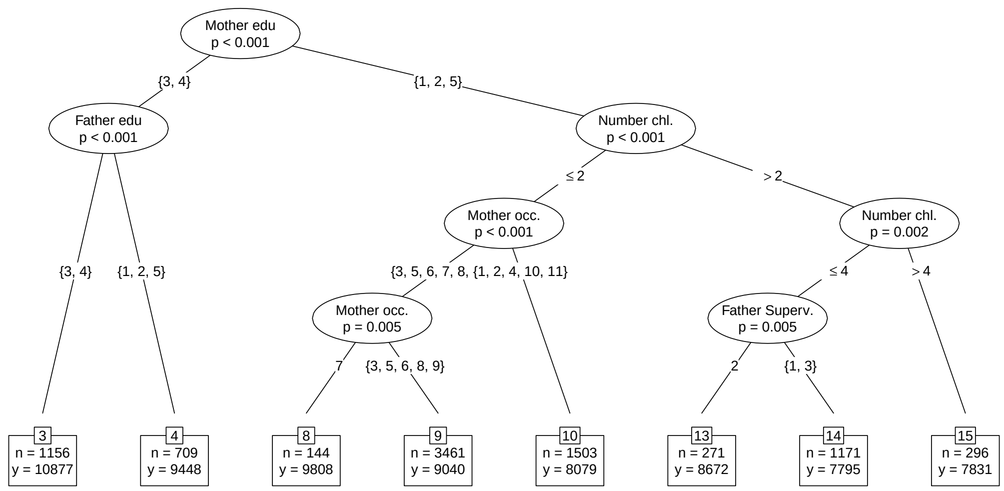
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.8 – Opportunity Tree (Cyprus)



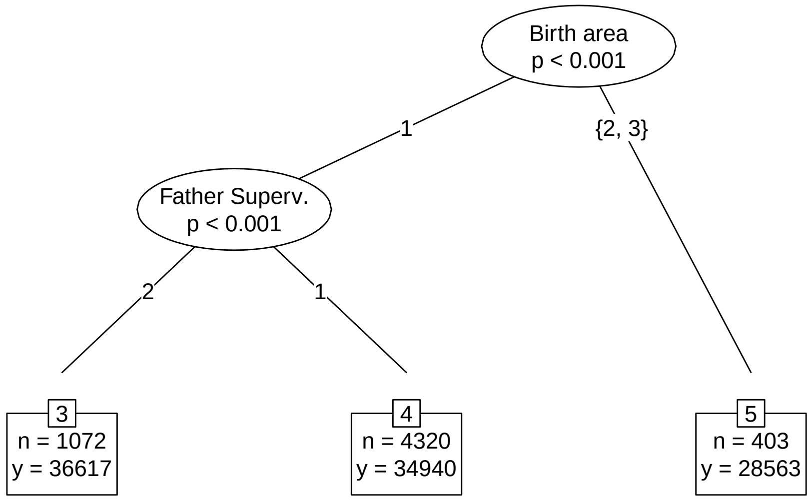
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.9 – Opportunity Tree (Czech Republic)



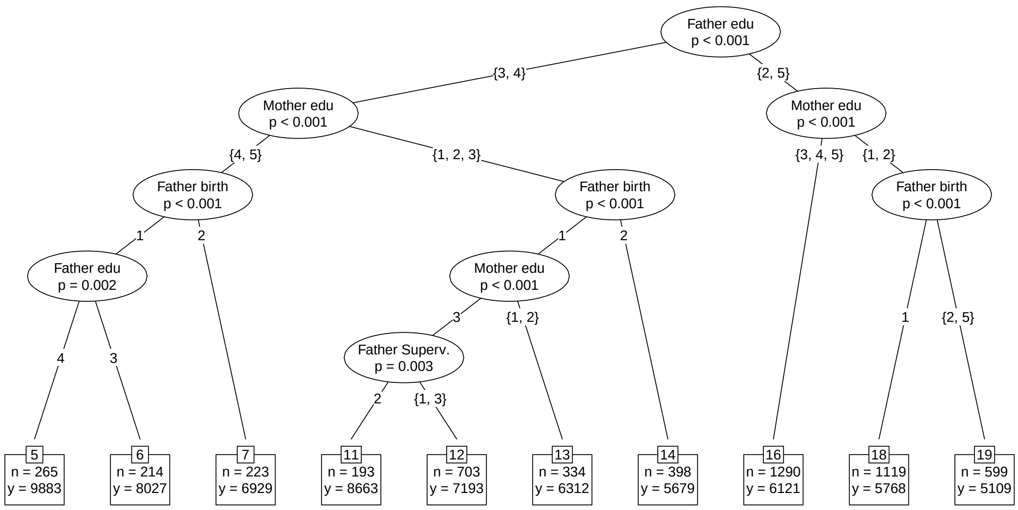
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.10 – Opportunity Tree (Denmark)



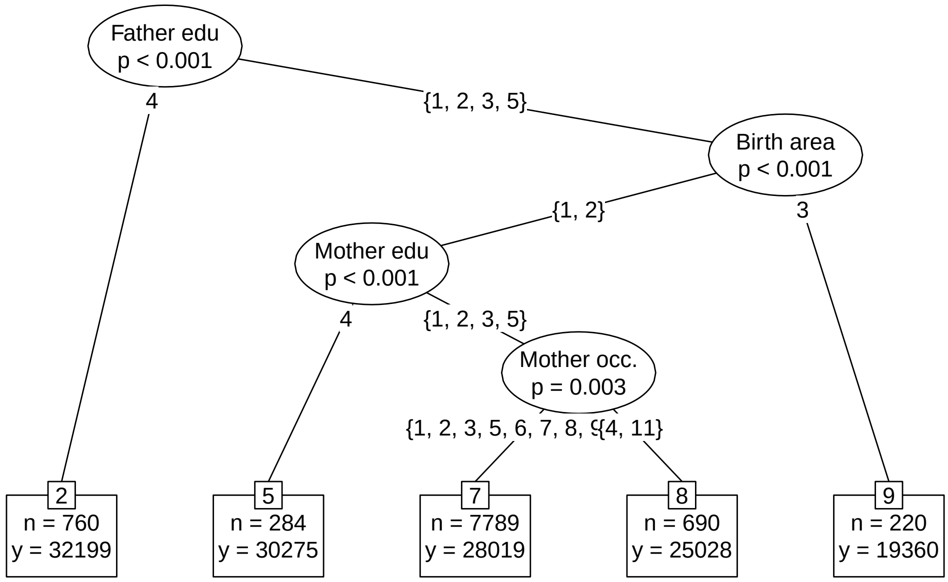
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.11 – Opportunity Tree (Estonia)



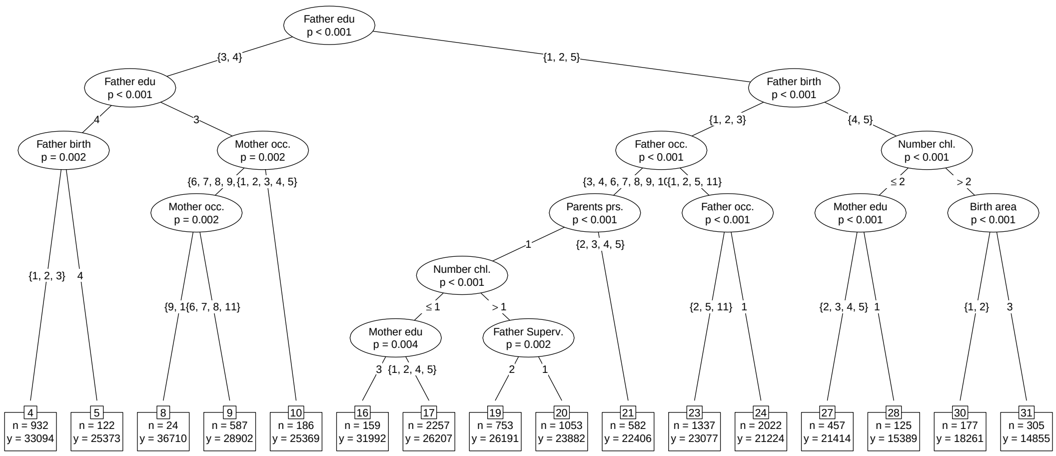
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.12 – Opportunity Tree (Finland)



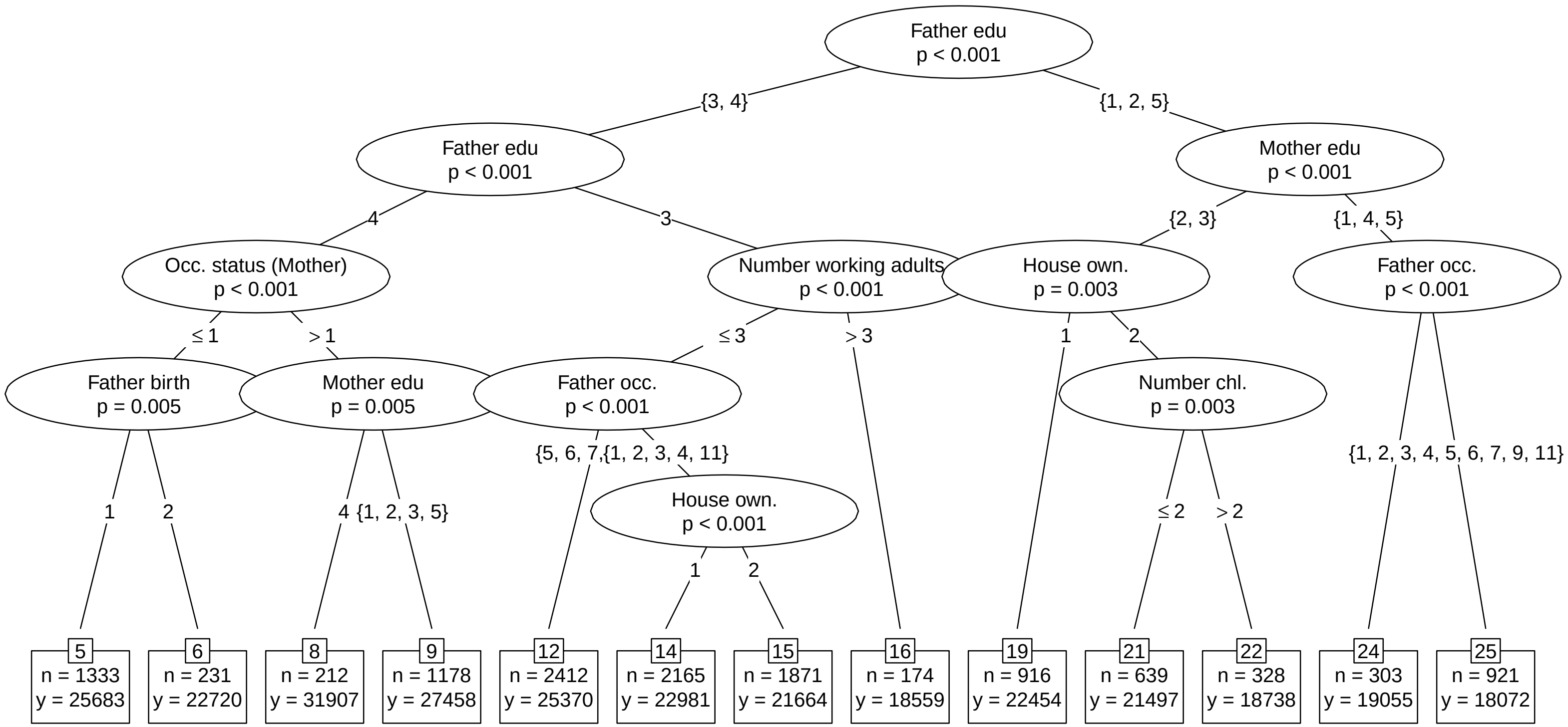
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the *p*-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.13 – Opportunity Tree (France)



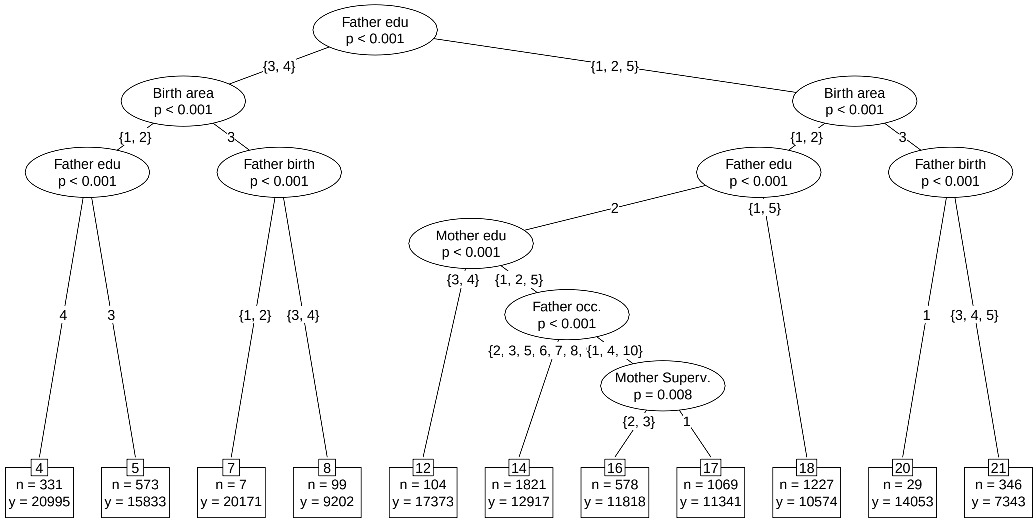
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the *p*-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.14 – Opportunity Tree (Germany)



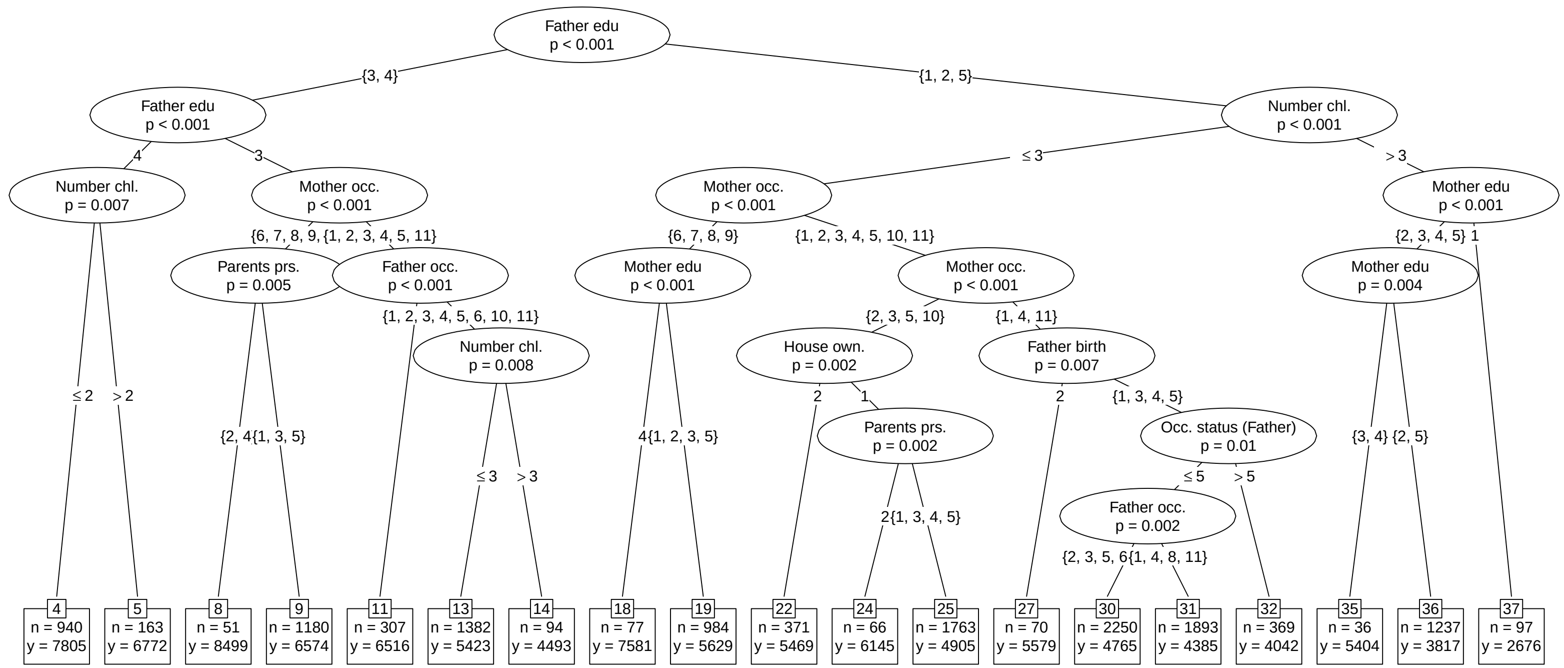
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.15 – Opportunity Tree (Greece)



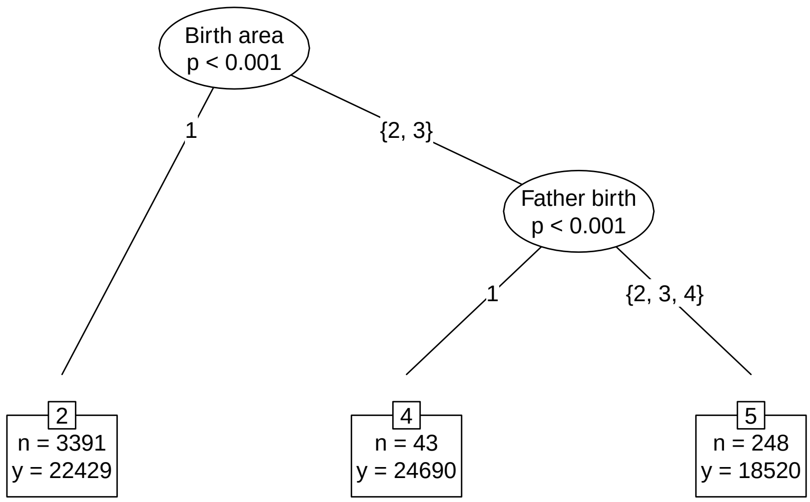
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.16 – Opportunity Tree (Hungary)



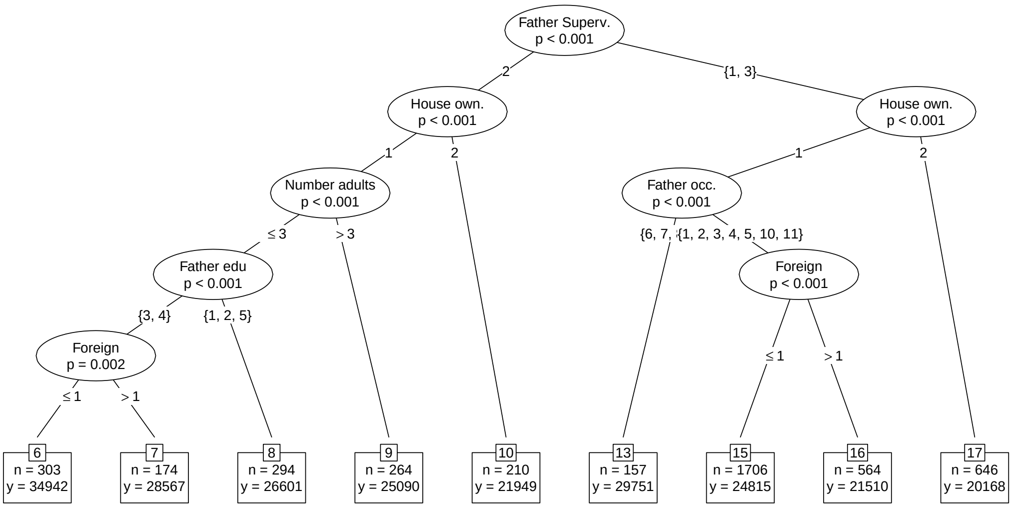
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.17 – Opportunity Tree (Iceland)



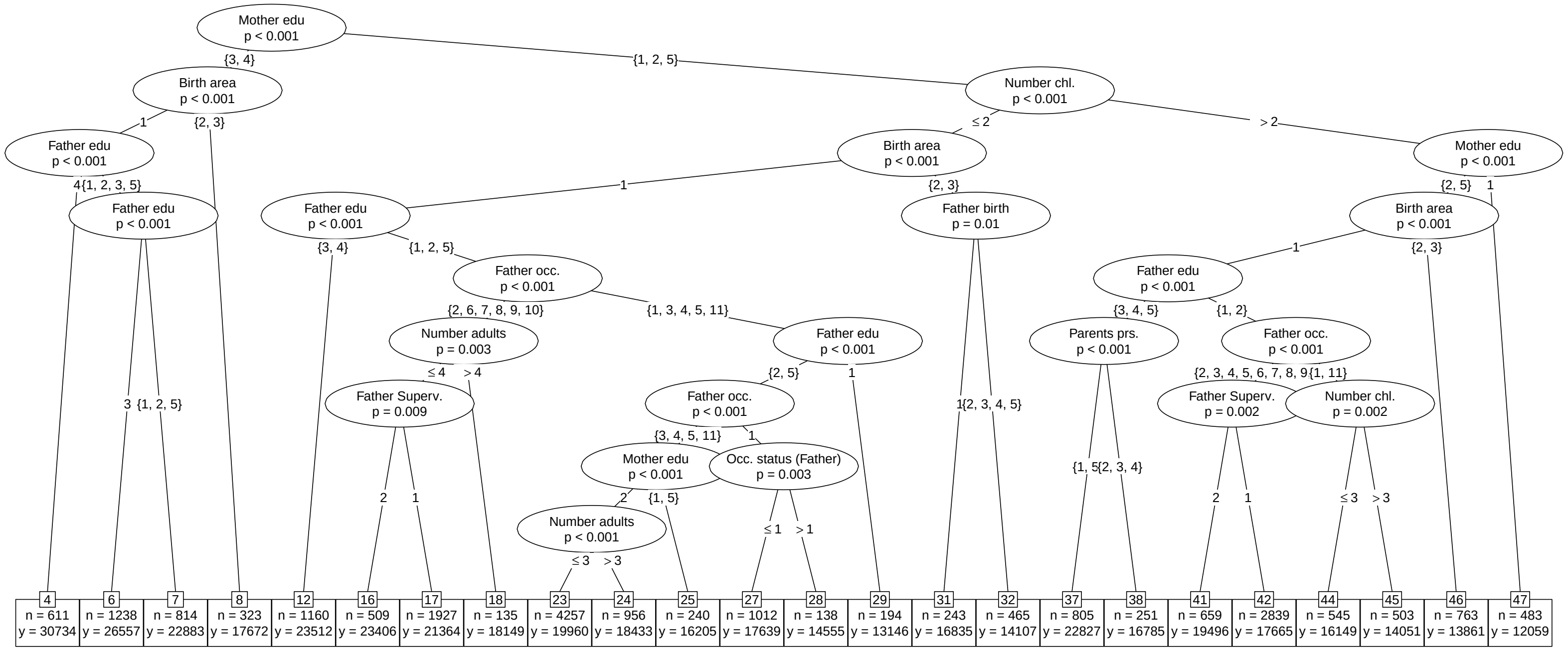
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.18 – Opportunity Tree (Ireland)



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

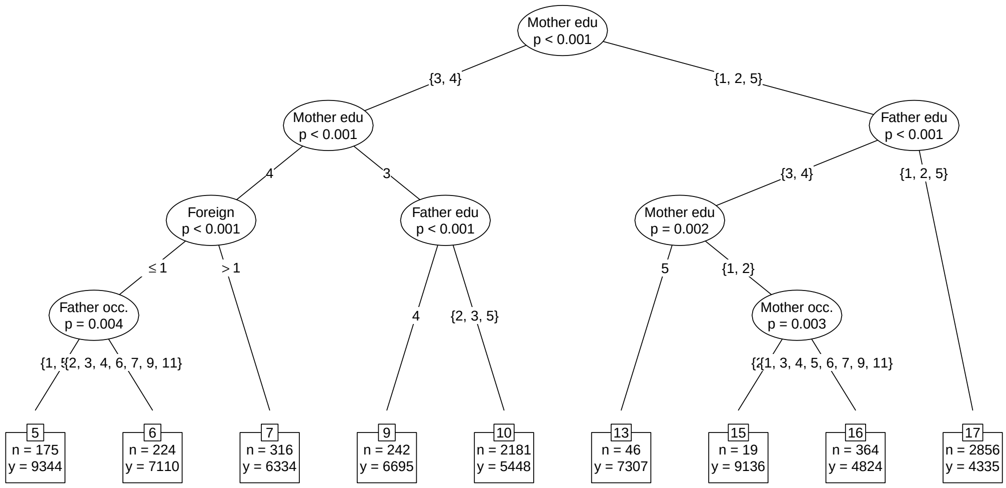
FIGURE S.19 – Opportunity Tree (Italy)



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).

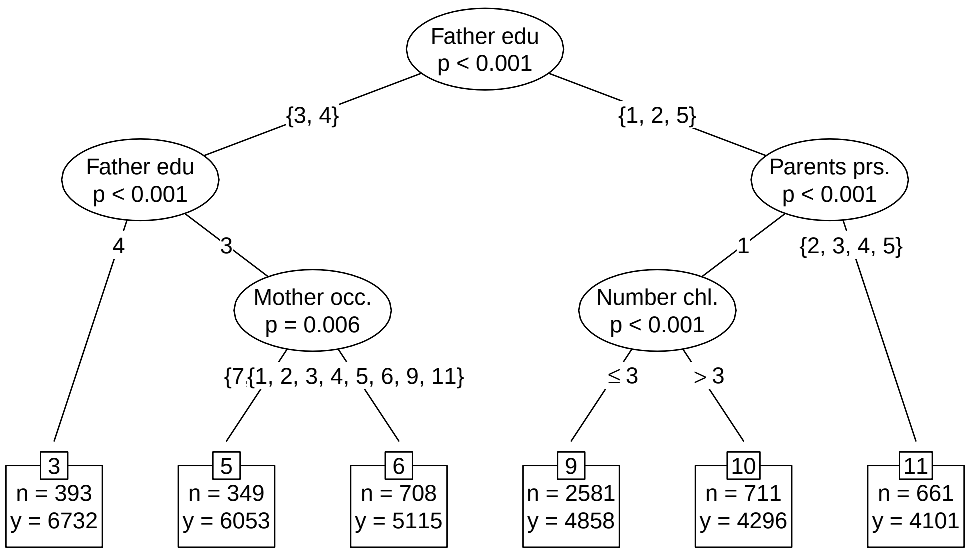
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.20 – Opportunity Tree (Latvia)



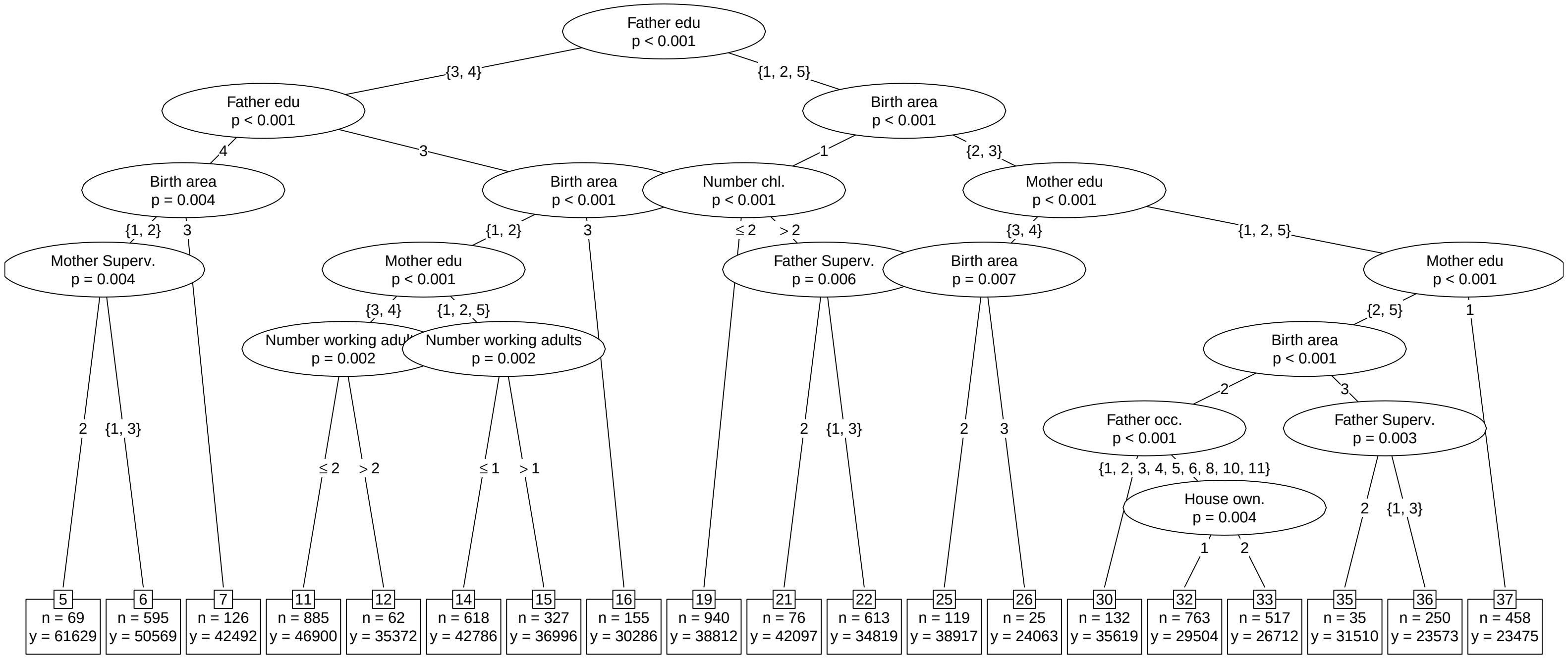
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.21 – Opportunity Tree (Lithuania)



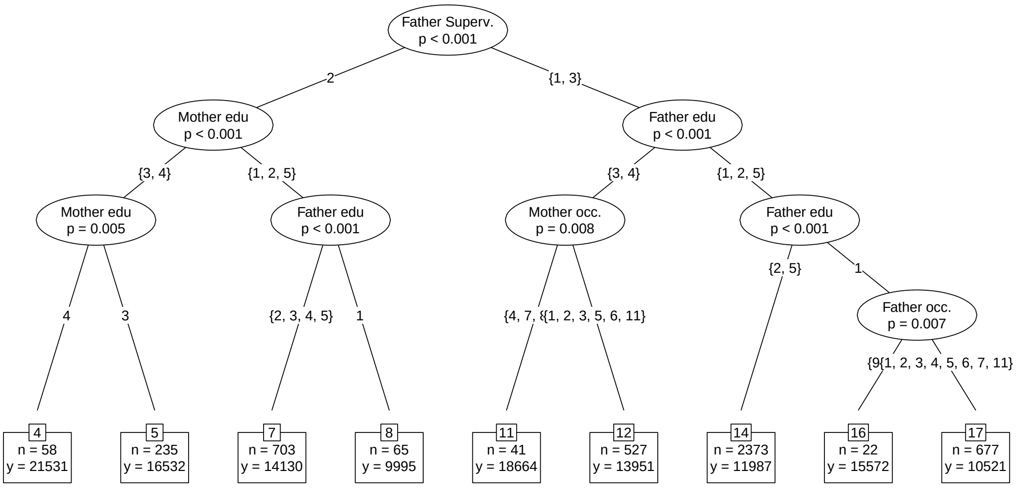
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.22 – Opportunity Tree (Luxembourg)



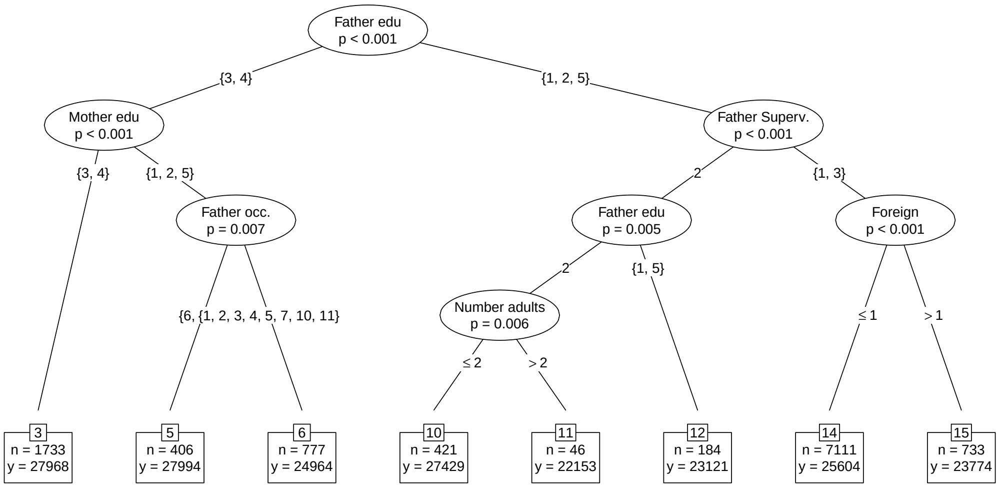
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.23 – Opportunity Tree (Malta)



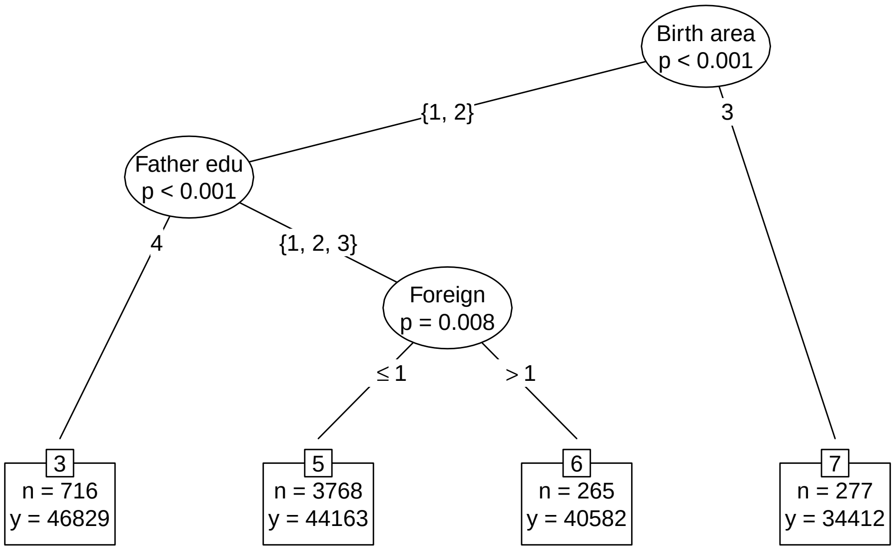
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.24 – Opportunity Tree (Netherlands)



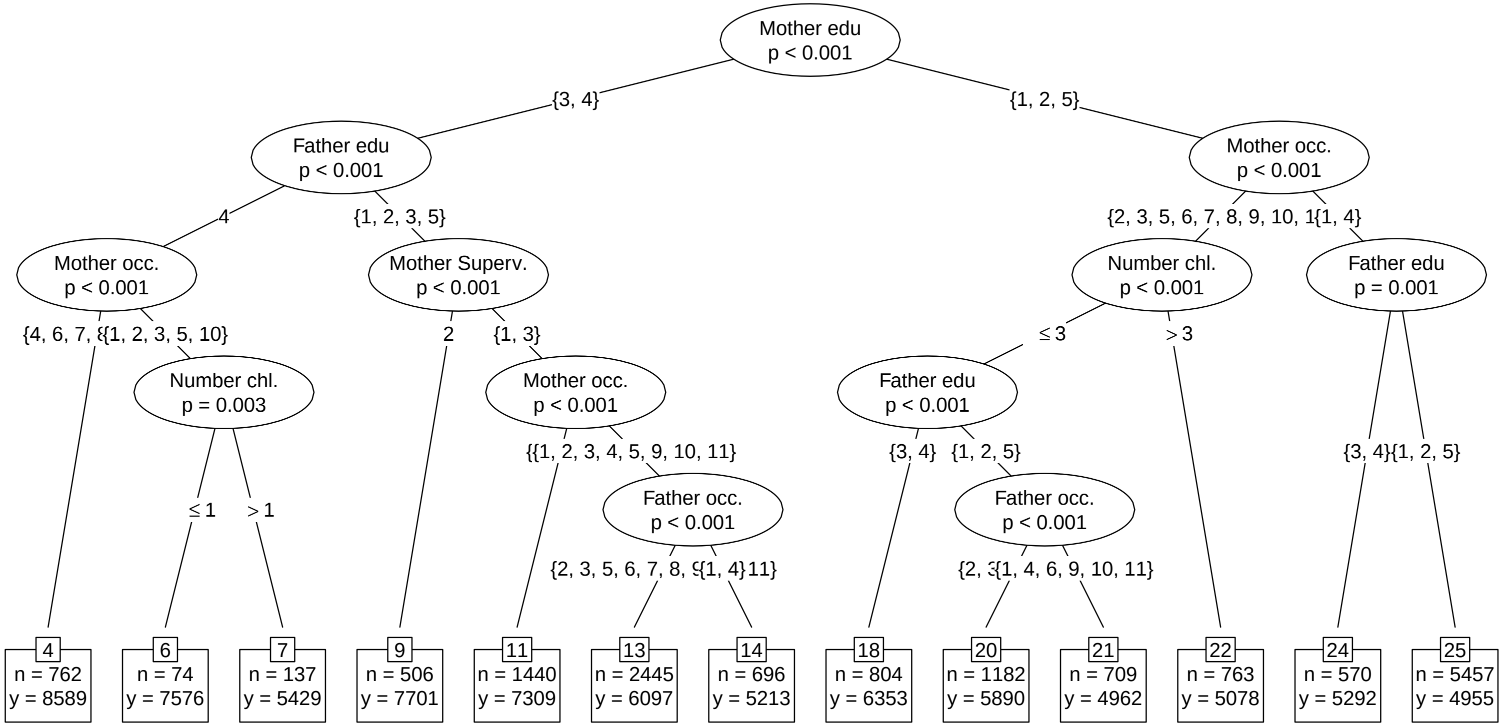
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.25 – Opportunity Tree (Norway)



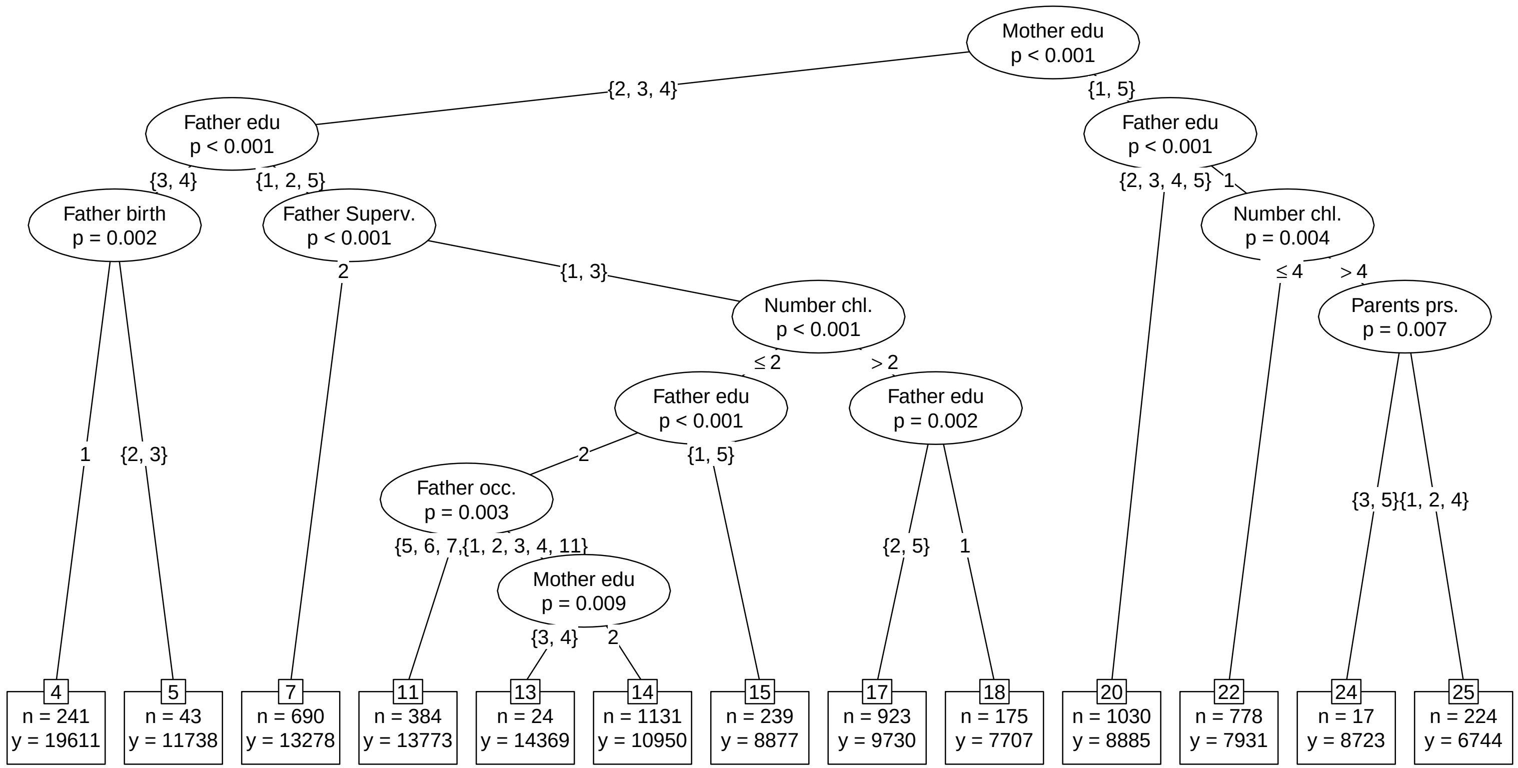
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.26 – Opportunity Tree (Poland)



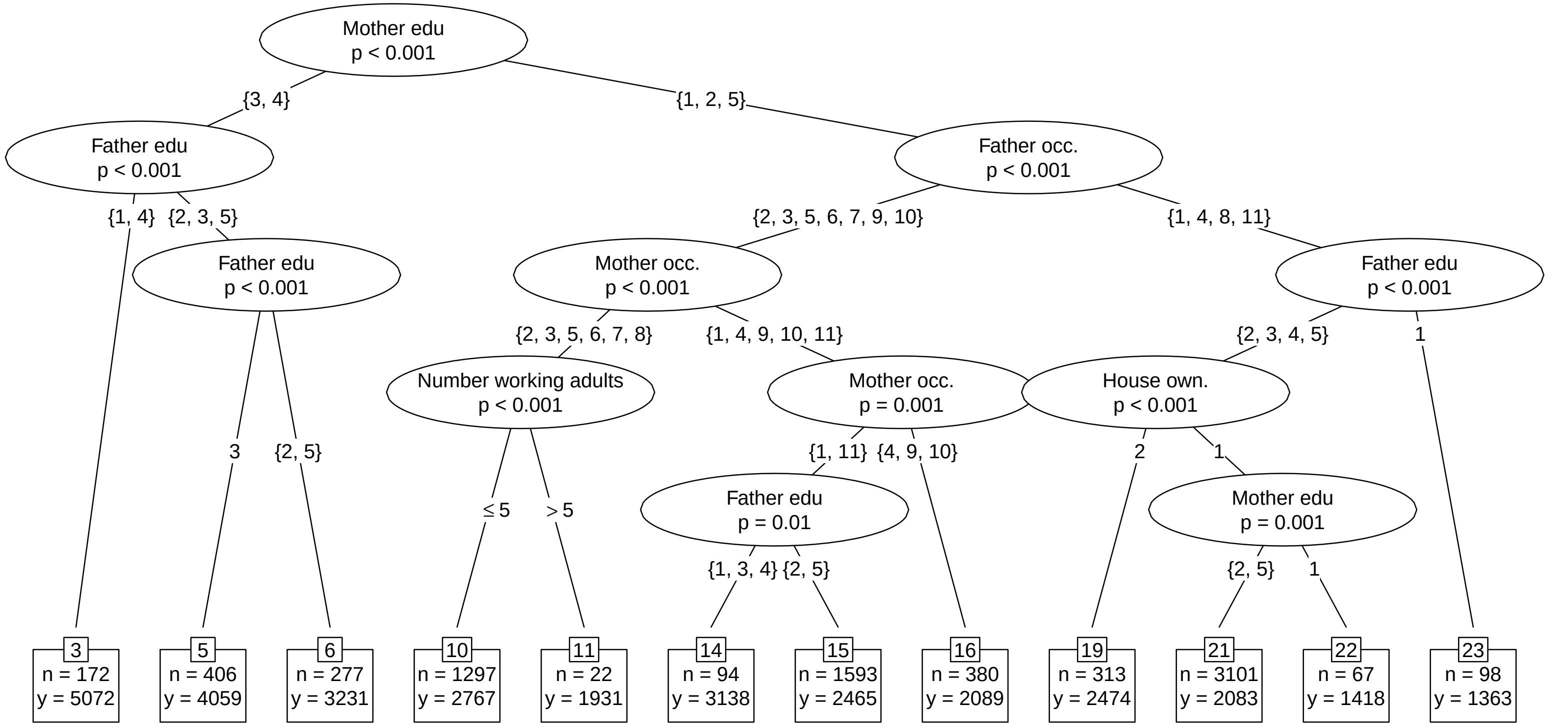
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the *p*-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.27 – Opportunity Tree (Portugal)



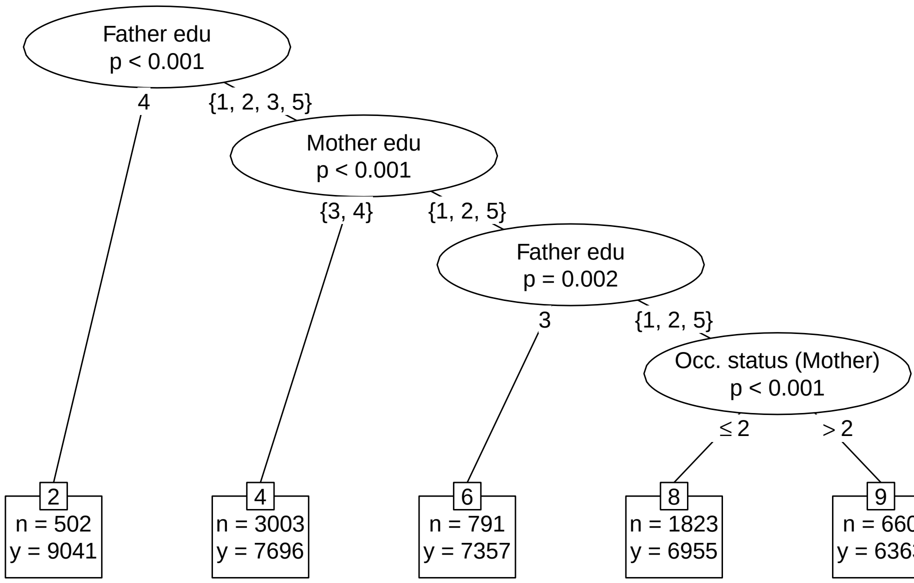
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the *p*-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.28 – Opportunity Tree (Romania)



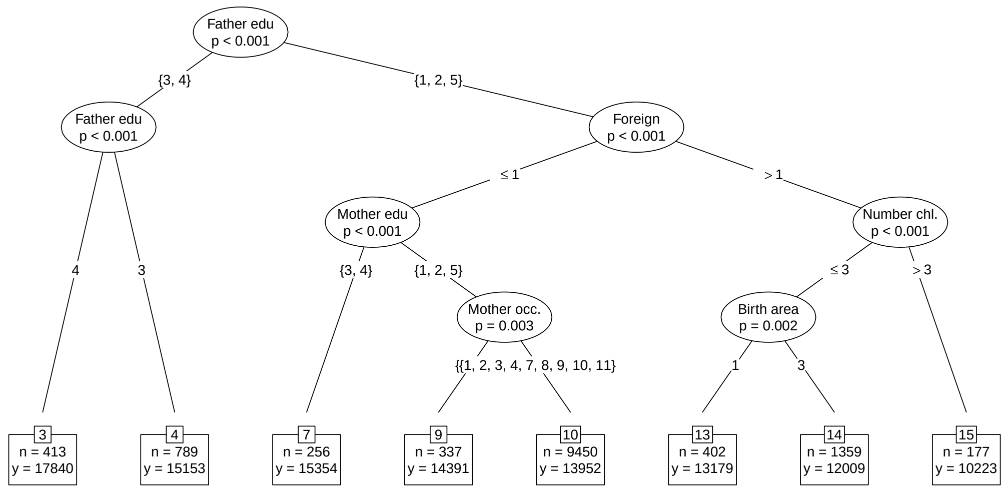
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.29 – Opportunity Tree (Slovakia)



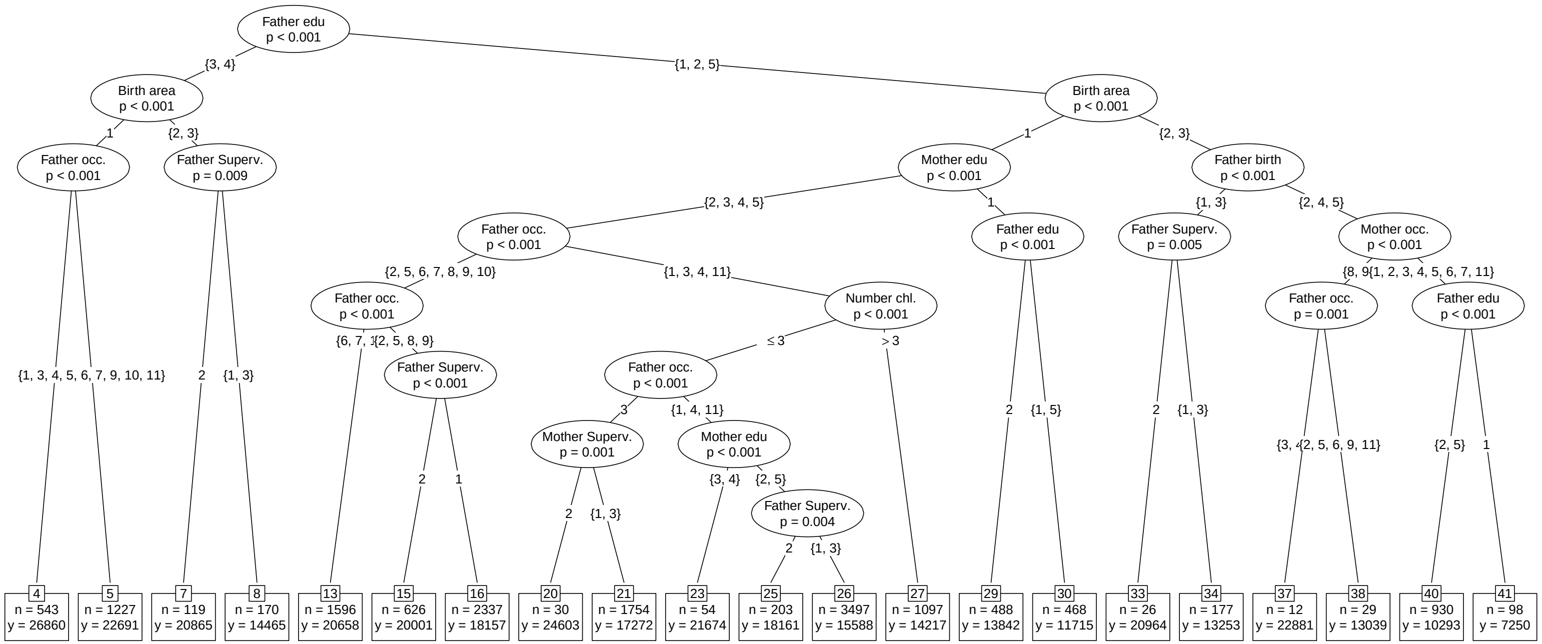
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.30 – Opportunity Tree (Slovenia)



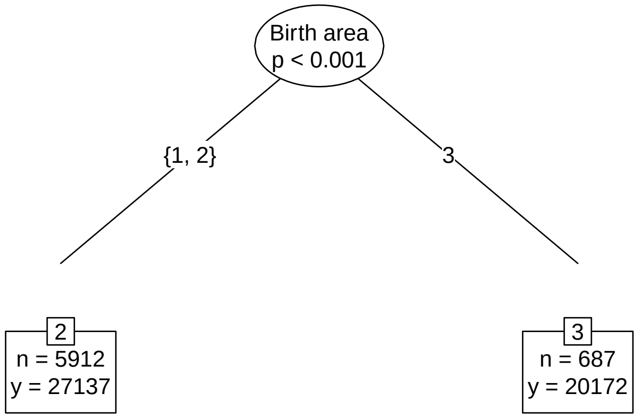
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.31 – Opportunity Tree (Spain)



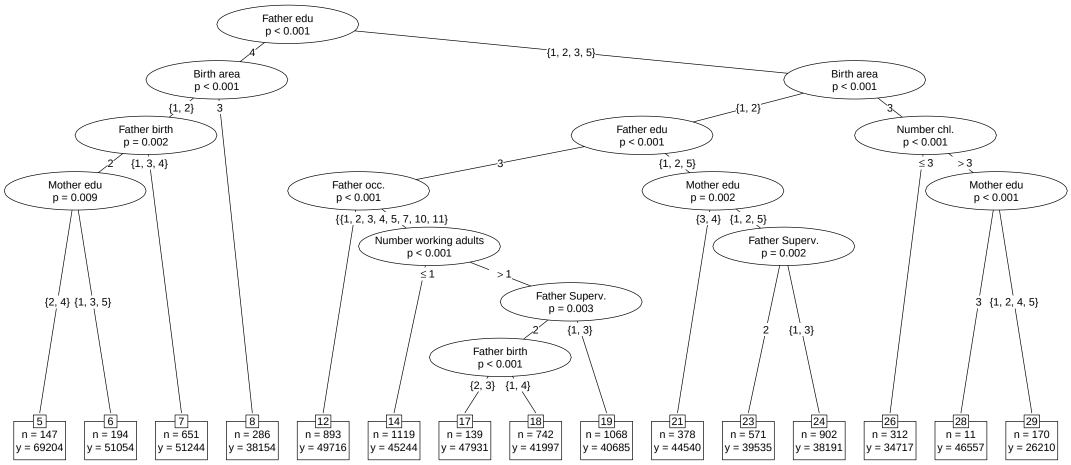
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the *p*-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.32 – Opportunity Tree (Sweden)



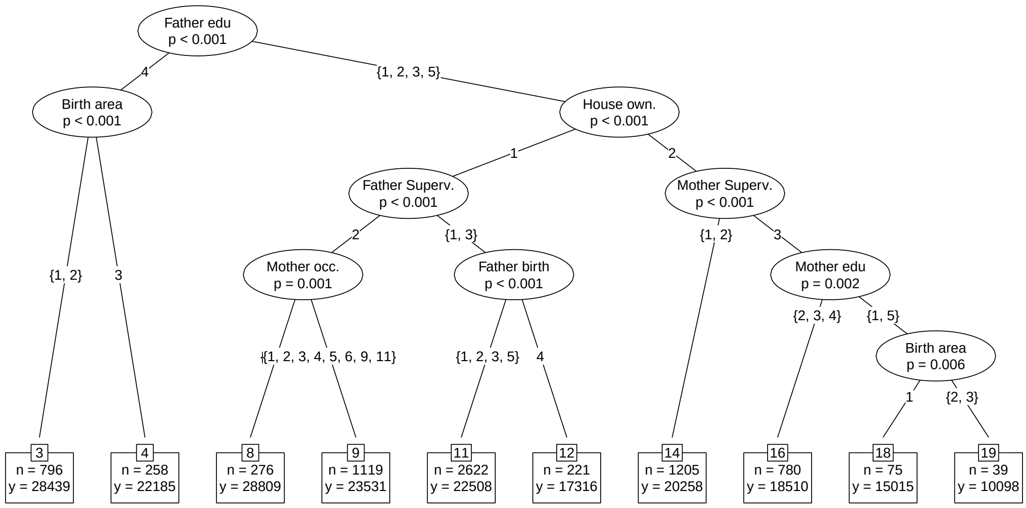
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.33 – Opportunity Tree (Switzerland)



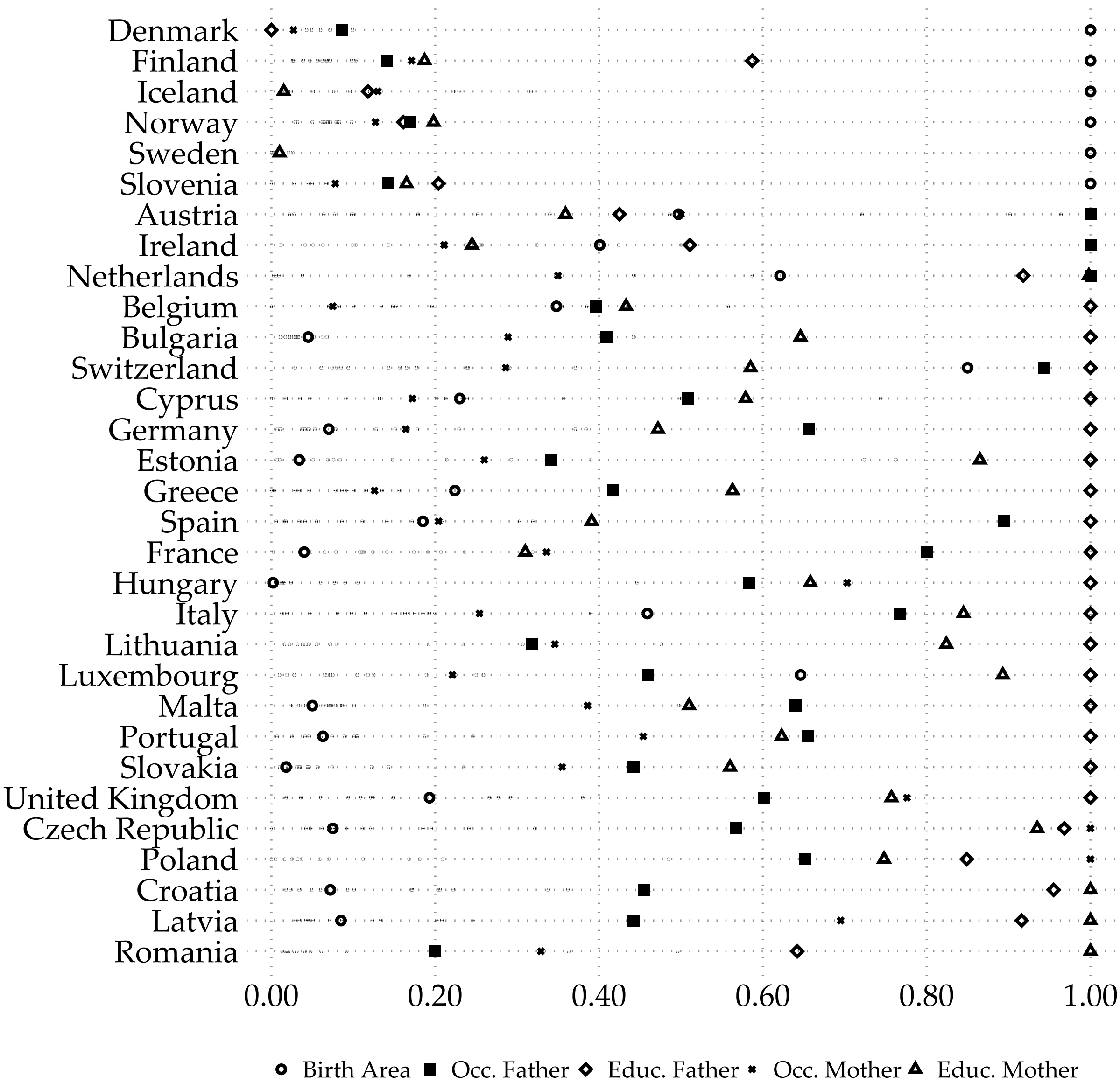
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.34 – Opportunity Tree (United Kingdom)



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section 3.1). The set of observed circumstances Ω used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while the rectangular boxes indicate terminal nodes. Within each ellipse we indicate the splitting variable as well as the p-value associated with the respective split. The first number inside the terminal nodes indicates the population share belonging to the circumstance type, while the second number shows the respective estimate of the conditional expectation y^C.

FIGURE S.35 – Variable Importance Plot from Forests



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: Each dot shows the importance of a particular circumstance variable ω^p. Variable importance is measured by the decrease in MSE^{OOB} after permuting ω^p such that it is orthogonal to y. The importance measure is standardized such that the circumstance with the greatest importance in each country equals 1.

References

BRUNORI, Paolo, Vito PERAGINE, and Laura SERLENGA (2019). "Upward and downward bias when measuring inequality of opportunity". *Social Choice and Welfare* 52, pp. 635–661.

FRIEDMAN, Jerome, Trevor HASTIE, and Robert TIBSHIRANI (2009). *The elements of statistical learning*. New York: Springer.