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

Paolo Brunori†, Paul Hufe‡, Daniel Gerszon Mahler§

June 25, 2020

Abstract

In this paper we propose the use of machine learning methods to estimate inequality of opportunity. We illustrate how our proposed methods – conditional inference regression trees and forests – represent a substantial improvement over existing estimation approaches. First, they reduce the risk of ad-hoc model selection. Second, they establish estimation models by trading off upward and downward bias in inequality of opportunity estimations. Finally, regression trees can be graphically represented; their structure is immediate to read and easy to understand. This makes the measurement of inequality of opportunity more easily comprehensible to a large audience. 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 researchers to overestimate (underestimate) inequality of opportunity by up to 300% (40%) in comparison to our preferred method.

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

*We would like to thank Chiara Binelli, Marc Fleurbaey, 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.

†University of Florence, paolo.brunori@unifi.it.

‡Corresponding author: University of Munich and ifo Institute, ifo Center for Macroeconomics and Surveys, Poschingerstr. 5, 81679 Munich, paul.hufe@econ.lmu.de.

§The World Bank, dmahler@worldbank.org.

INTRODUCTION

Equality of opportunity is an important ideal of distributive justice. It has widespread support in the general public (Alesina et al., 2018; Cappelen et al., 2007) and its realization has been identified as an important goal of public policy intervention (Chetty et al., 2016; Corak, 2013). 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 approaches to estimate inequality of opportunity 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. For example, in comparison to our preferred method, current estimation approaches overestimate inequality of opportunity in Scandinavian countries by close to 300%. While these figures may inform policy debates about inclusive institutions, they are the result of overfitted estimation models that fail to replicate in independent samples of the same underlying population. This example illustrates that the choice of appropriate model specifications is of great importance for the analysis of institutional configurations and the ensuing policy debate.

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

[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).

on this distinction, inequality of opportunity measures quantify the extent to which individual outcomes are determined by circumstance characteristics. In particular, inequality of opportunity is frequently measured by using a set of circumstances to predict an outcome of interest and calculating inequality in the predicted outcomes: the more predicted outcomes diverge, the more circumstances beyond individual control influence outcomes, and the more inequality of opportunity there is.

Estimates of inequality of opportunity matter for development policy in several ways. They identify the most opportunity deprived and can thus help to improve the targeting of resources through social protection programs and other transfers (Belhaj Hassine, 2011). Comparing estimates over time and across countries may help to determine the extent to which a policy area should be prioritized (Ferreira et al., 2018). Given that inequality of opportunity is highly associated with both inequality of outcomes and pro-poor growth, policies to reduce inequality of opportunity are likely to reduce inequality and poverty as well (Marrero et al., 2016; Peragine et al., 2013).

In spite of their policy relevance, 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 the increasing availability of high-quality datasets that provide very detailed information with respect to individual circumstances (Björklund et al., 2012a; 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 the functional form according to which circumstances co-produce the outcome of interest. For example, it is a well-established finding that the influence of similar child care arrangements on various life outcomes varies strongly by biological sex (Felfe and Lalive, 2018; García et al., 2018). In contrast to such evidence, many

[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.

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 another 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 and popularized by Breiman et al. (1984), Breiman (2001), and Morgan and Sonquist (1963), regression trees and forests belong to a set of machine learning methods that is increasingly integrated into the statistical toolkit of economists (Athey, 2018; Mullainathan and Spiess, 2017; Varian, 2014). By drawing on a clear-cut algorithm, they obtain predictions without assumptions about which and how circumstances interact in shaping individual opportunities. Hence, the model specification is no longer a judgment call of the researcher but an outcome of 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 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.

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 Scandinavian countries by close to 300%, whereas they underestimate the extent of inequality of opportunity in Germany by more than 40%. Hence, cross-country comparisons based on standard estimation approaches yield misleading recommendations with respect to the need for policy intervention in different societies. We illustrate how regression trees and forests can be used to analyze opportunity structures in different societies. We find that mothers' education and occupation are the most important predictors of children's income in Eastern Europe, while in Western/Southern Europe fathers' occupation and education are most important, and in Northern Europe area of birth is most important. Although we are careful to highlight the non-causal nature of our estimates, such analyses provide useful starting points for policymakers to target areas for opportunity equalizing reforms.

In a parallel paper, Blundell and Risa (2019) apply machine learning methods to the estimation of intergenerational mobility – a literature in which similar issues of model selection arise.[3] In particular, they use machine learning methods to validate rank-rank estimates of intergenerational mobility against an extended set of child circumstances to assess the completeness of the prevalent intergenerational mobility approach as a measure for equal opportunities. In contrast to their work, we directly estimate inequality of opportunity statistics. As a consequence, our focus is less on the downward bias that follows from focusing on one circumstance characteristic only (e.g. 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.

The remainder of this paper is organized as follows: section I gives a brief introduction to current empirical approaches in the literature on inequality of opportunity. Section II introduces conditional inference regression trees and forests, and illustrates how to use

[3]These issues include the influence of non-linearities along the parental distribution (Björklund et al., 2012b; Corak and Piraino, 2011) and the question of whether intergenerational persistence is sufficiently characterized by focusing on the parent-child link only (Braun and Stuhler, 2018; Mare, 2011). Furthermore, recent works in this branch of the literature go beyond single indicator models and use many proxy variables to construct comprehensive indicators for the underlying parental social status (Vosters, 2018; Vosters and Nybom, 2018).

them in the context of inequality of opportunity estimations. An empirical illustration based on the EU Survey of Income and Living Conditions is contained in section III. 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. Lastly, section IV concludes.

I EMPIRICAL APPROACHES TO EQUALITY OF OPPORTUNITY

Theoretical Set-up and Notation. Consider a population $\mathbf{N} := \{1, ..., N\}$ and an associated vector of non-negative outcomes $y = (y_1, ..., y_N)$. Outcomes are the result of two sets of factors: First, a set of *circumstances* beyond individual control: $\mathbf{\Omega} := C^1 \times ... \times C^P$. Second, a set of *efforts* $\mathbf{\Theta} := E^1 \times ... \times E^Q$. In what follows, $\mathbf{\Omega}$ and $\mathbf{\Theta}$ will be referred to as the circumstance and effort space, spanned by the dimensions $(C^p,\ p = 1, ..., P)$ and $(E^q,\ q = 1, ..., Q)$, respectively. We define the $(P \times 1)$-vector $\omega_i \in \mathbf{\Omega}$ as a comprehensive description of the circumstances with which $i \in \mathbf{N}$ is endowed. Analogously we define the $(Q \times 1)$-vector $\theta_i \in \mathbf{\Theta}$ as a comprehensive description of the efforts that are exerted by $i \in \mathbf{N}$.

The outcome generating function can be defined as follows:

$$\mathbf{\Omega} \times \mathbf{\Theta} \ni (\omega, \theta) \mapsto d(\omega, \theta) =: y,\ \ y \in \mathbb{R}_+, \tag{1}$$

such that for every $i \in \mathbf{N}$, the individual outcome y_i is a function of her circumstances ω_i and the effort θ_i she exerts. Individual effort exertion is plausibly co-determined by circumstance characteristics. We follow Roemer (1998) in adopting a relative conception of effort. Normatively, this assumption entails a stance according to which outcome differences due to a correlation between circumstances and effort constitute a violation of the opportunity egalitarian ideal. For example, if individuals work shorter hours due to wage discrimination in the labor market we would deem the ensuing income differences worth of compensation. Econometrically, this assumption entails that θ is purged of its correlation with circumstance characteristics ω such that effort is independently distributed

of circumstance characteristics (see Lefranc et al., 2009; Roemer and Trannoy, 2015, for discussions). While such a conception is in line with the majority of the literature, our estimation approach is not dependent on it and can be easily extended to alternative cuts between ω and θ (Jusot et al., 2013).

Based on the realizations of individual circumstances ω_i the population can be partitioned into a set of *types*. We define the type partition $\mathbf{T} = \{t_1, ..., t_M\}$, such that individuals are member of one type if they share the same set of circumstances: $i, j \in t_m \Leftrightarrow \omega_i = \omega_j, \ \forall t_m \in \mathbf{T}, \ \forall i, j \in \mathbf{N}$. Hence, types define one particular way of partitioning the population into groups, where group membership indicates uniformity in circumstances.

Measurement. Opportunity egalitarians are averse to inequality to the extent that it is rooted in circumstance factors that are beyond individual control. They are agnostic towards inequalities that originate from differences in 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. Each of these different formulations is pinned down by combining a *principle of compensation* with a *principle of reward* (Aaberge et al., 2011; Almås et al., 2011; Fleurbaey, 1995; Ramos and Van de gaer, 2016). The former specifies how differences due to circumstances should be compensated. The latter specifies to what extent differences due to effort should be respected. In this work we exclusively focus on the principles of *ex-ante compensation* and *utilitarian reward*. Measures satisfying these two principles were first proposed in Checchi and Peragine (2010) and Van de gaer (1993). They are the most widely applied formulations in empirical works on equality of opportunity. To keep our analysis tractable we restrict ourselves to this particular conception of inequality of opportunity. However, our estimation approach is not dependent on it and can be easily extended to alternative measures of inequality of opportunity.

The ex-ante view of compensation focuses on between-type differences in the value of opportunity sets without paying attention to the specific effort realizations of individual type members. That means, we always prefer a distribution y' over y if the former is obtained from the latter by making a transfer from a more advantaged type to a less advantaged type. Utilitarian reward specifies zero inequality aversion with respect to

income differences within a type. As a consequence, the value of the opportunity set of a type 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).[4] 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) would value compensating transfers to the most disadvantaged types more than the Gini index. In this work we are agnostic towards the normatively correct choice of $I()$. While we will present our main results in terms of the Gini index, we provide robustness checks based on other inequality indexes in Supplementary Material B.

Note that the measurement of inequality of opportunity can also be understood as a decomposition exercise where total inequality is split into a between- and a within-group

[4]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} equals 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 of equal income.

component. It thereby relates to the broad literature on distributional decompositions in labor economics (Fortin et al., 2011). However, it is important to highlight that opportunity egalitarians view differences among circumstance groups as normatively objectionable regardless of whether these differences are the result of compositional differences in (un)observed characteristics (e.g. educational achievement and occupational choices) or the return to such characteristics. While distinctions among these different explanations are important for the design of appropriate policy responses, they are of indifference for the measurement of inequality of opportunity in the ex-ante utilitarian sense.

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

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

where $\mathbb{E}(y|\omega)$ captures variation due to observed circumstances. The iid error term ϵ captures variation due to unobserved circumstances and individual effort. The fact that ϵ representes both fair (individual effort) and unfair (unobserved circumstances) determinants of individual outcomes illustrates that the 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? This task is complicated by the fact that 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 of y^C from the predicted values:

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

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

where $\omega_i^p \in \mathbf{\Omega}$. The literature refers to this estimation procedure as the *parametric approach* (Bourguignon et al., 2007; Ferreira and Gignoux, 2011).

Another common estimator of y^C comes from an approach where the researcher partitions the sample into mutually exclusive types based on the realizations of all circumstances under consideration. An estimate of y^C is then constructed from the average outcome values within types:

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

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

Both approaches face empirical challenges which 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 space to obtain statistically meaningful estimates of the relevant parameters. Assume for example, that the researcher observes ten circumstance variables with three expressions each – a quantity easily observed in many household surveys. Implementing the non-parametric approach would require the estimation of $3^{10} = 59,049$ group means which is hardly feasible given the sample sizes of most household surveys. The necessary process of restricting the circumstance space again remains at the researcher's discretion.

The previous discussion illustrates that common approaches leave the researcher to her own devices when it comes to 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, t_m. For each possible number of latent types, M, 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 cannot solve the problem of model selection once the potential number of type characteristics exceeds the available degrees of freedom. In such cases, the latent class approach replicates the limitations of the parametric and the non-parametric approach: the researcher must pre-select the relevant set of circumstances, their subpartition, and the respective interactions. Furthermore, 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 not well-suited for capturing the dependence between circumstances and a particular outcome of interest.

In the following we will show how the outlined shortcomings of existing approaches can be addressed by regression trees and forests.

II 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.

While there are many supervised learning methods to solve prediction problems, trees and forests are particularly attractive in our setting since they are very flexible in accounting for non-linearities and effective in excluding features that are unrelated to the outcome of interest (Athey and Imbens, 2019). Moreover, in the context of inequality of opportunity estimations they strike a balance between prediction accuracy and inter-

pretability.[5]

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. Furthermore, their simple graphical illustration is particularly instructive for longitudinal or cross-sectional comparisons of opportunity structures. 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.

Conditional Inference Trees

Tree-based methods obtain predictions for outcome y as a function of the input variables $x = (x^1, ..., x^k)$. Specifically, they use the sample $\mathcal{S} = \{(y_i, x_i)\}_{i=1}^{S}$ to divide the population into non-overlapping groups, $\mathbf{G} = \{g_1, ..., g_m, ..., g_M\}$, where each group g_m is homogeneous in the expression of some input variables. These groups are called *terminal nodes* or *leafs* in a regression tree context. The conditional expectation for observation i is estimated from the mean outcome $\hat{\mu}_m$ of the group g_m to which the i^{th} observation is assigned. Hence, in addition to the observed outcome vector $y = (y_1, ..., y_i, ..., y_N)$ one obtains a vector of predicted values $\hat{y} = (\hat{f}(x_1), ..., \hat{f}(x_i), ..., \hat{f}(x_N))$, where

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

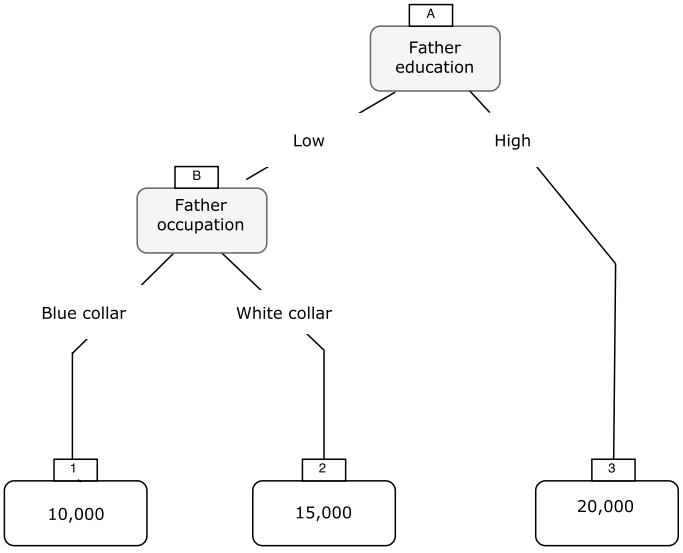
and N_m is the size of each group.

The mapping from regression trees to equality of opportunity estimation is straightforward. Conditional on the input variables being circumstances only, each resulting group $g_m \in \mathbf{G}$ can be interpreted as a circumstance type $t_m \in \mathbf{T}$. Furthermore, $\hat{y}$ is analogous to an estimate of the counterfactual distribution y^C which in turn can be used for the construction of ex-ante utilitarian measures of inequality of opportunity.

[5]Furthermore, it has been shown that many other ensemble methods show negligible gains in prediction accuracy in comparison to random forests when applied to social science questions, e.g. see Blundell and Risa (2019) for an example.

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 $\omega^p \in \mathbf{\Omega}$. 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 $\omega^q \in \mathbf{\Omega}$), and so on. Graphically, this division into groups can be presented like an upside-down tree (Figure 1).

Figure 1: Exemplary Tree Representation



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

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 procedure:[6] First, they test the relationship between the outcome variable and each circumstance variable in a univariate way. The circumstance that is most related to the outcome is chosen as the potential splitting variable. Second, if the dependence between the outcome and the splitting variable is sufficiently strong,

[6]See Appendix A.I for a precise statement of the algorithm.

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. Third, 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 structure and depth of the resulting opportunity tree hinges crucially on the level of α^*, i.e. the critical value for the permutation tests to reject the null hypothesis. 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 Appendix Figure A.1.

Conditional Inference Forests

Regression trees solve the model selection problem outlined in section I and provide a simple and standardized way of dividing the population into types. However, constructing estimates for the counterfactual distribution y^C from conditional inference trees suffers from three shortcomings: first, the structure of trees – and therefore the estimate of the relevant distribution y^C – is fairly sensitive to alternations in the respective 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. On the one hand, this is fully consistent with Roemer's theory by which circumstances partition the population into types. On the other hand, the best model for constructing $\hat{y}$ may in fact be linear in some circumstances. Third, trees make only limited use of the information inherent in the set of observed circumstances since some of the circumstances $\omega^p \in \mathbf{\Omega}$ 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 a below α^*.

In what follows we will introduce conditional inference forests (Biau and Scornet, 2016;

Breiman, 2001) which address all three of these shortcomings.

Forest Construction. Random forests create many trees and average over all of these when making predictions. Trees are constructed according to the same 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 case, we randomly select half of the observations for each tree, and estimate B^* such trees in total. Second, only a random subset of circumstances $\{\omega^p \in \mathbf{\Omega} : p \in \bar{\mathbf{P}} \subset \{1, ..., P\}\}$ of size $\bar{P}^*$ is allowed to be used at each splitting point. Together these two tweaks remedy the shortcomings of single conditional inference trees. First, averaging over the B^* predictions cushions the variance in the estimates of y^C and smoothes the non-linear impact of circumstance characteristics. Second, drawing only on subsets of all circumstance variables increases the likelihood that all observed circumstances with informational content will be identified as the 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 to be 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 the empirical procedures are disclosed in Appendix A.II.

III EMPIRICAL APPLICATION

In this section we provide an illustration of the machine learning approach using harmonized survey data from 31 European countries. We will 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 will be made

along two dimensions.

First, we evaluate the different estimation approaches by comparing their out-of-sample mean squared error (MSE). The MSE provides a standard statistic to evaluate the prediction quality of different models by representing the variance-bias trade-off. In the context of constructing an estimate of the conditional income distribution y^C, this property is equivalent to trading-off upward and downward biases in inequality of opportunity estimates: 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. A thorough illustration of this mapping is provided in Appendix A.III.

Second, we compare the inequality of opportunity estimates emanating from the set of benchmark methods to the ones from regression trees and forests.

In a last step, we illustrate how regression trees and forests can be used to analyze opportunity structures in the population of interest.

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 income, poverty, and living conditions on an annual basis and covers a cross-section of 31 European countries in the 2011 wave.[7] For each country, EU-SILC provides a random sample of all resident, private households. The data is collected by the various national statistical agencies following common variable definitions and data collection procedures. It provides the official reference source for comparative statistics on income distribution and social inclusion in the European Union (EU) and therefore provides a degree of harmonization that makes it particularly suitable for methodological comparisons. We draw on the 2011 wave since it contains an ad-hoc module about

[7]The sample consists of Austria (AT), Belgium (BE), Bulgaria (BG), Switzerland (CH), Cyprus (CY), Czech Republic (CZ), Germany (DE), Denmark (DK), Estonia (EE), Greece (EL), Spain (ES), Finland (FI), France (FR), Croatia (HR), Hungary (HU), Ireland (IE), Iceland (IS), Italy (IT), Malta (MT), Lithuania (LT), Luxembourg (LU), Latvia (LV), Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Romania (RO), Sweden (SE), Slovenia (SI), Slovak Republic (SK), and Great Britain (UK).

the intergenerational transmission of (dis)advantages which allows us to construct finely-grained circumstance type partitions. The space of observed circumstances $\mathbf{\Omega}$ and their respective expressions are listed in Table 1. The list includes all variables of EU-SILC containing information about the respondent's characteristics at birth and their living conditions during childhood. Descriptive statistics concerning circumstances are reported in Supplementary Material A.

The unit of observation is the individual and the outcome of interest is equivalized disposable household income. The latter is obtained by dividing household disposable income by 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. Aware that inequality statistics tend to be heavily influenced by outliers (Cowell and Victoria-Feser, 1996) 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 59. To assure the representativeness of our country samples all results are calculated by using appropriate individual cross-sectional weights.

Table 2 shows considerable heterogeneity in the income distributions of the European country sample. While the average households in Norway (NO) and Switzerland (CH) obtained incomes above €40,000 in 2010, the average household income in Bulgaria (BG), Romania (RO) and Lithuania (LT) did not exceed the €5,000 mark. The lowest inequality prevails in the Nordic countries of Norway (NO), Sweden (SE) and Iceland (IS), all of which have Gini coefficients below 0.220. At the other end of the spectrum we find the Eastern European countries of Latvia (LV), Lithuania (LT) and Romania (RO) with Gini coefficients well above 0.330.

Table 2 also shows the sample size for each country. These figures include observations with missing values in one or more of the circumstances we use. The parametric approach, the non-parametric approach, and latent class analysis handle missing values by listwise

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 (including also armed force) (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 occupation (I-00)
    - Father/mother did not work, was unknown or was dead
11. Managerial position of the 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: Questions marked with * refer to the period when the respondent was approximately 14 years old. Item 11 is missing for Finland. We exclude subjective questions about the financial situation and the level of deprivation of the household of origin from the list of circumstances.

deletion. In contrast, conditional inference trees and forests make use of the full sample by allowing for surrogate splits. For each splitting point $\tilde{\omega}^*$, the algorithm searches for an alternative splitting point $\tilde{\omega}^+$ that mimicks the sample partition of $\tilde{\omega}^*$ to the greatest extent. All observations that lack information on $\tilde{\omega}^*$ are then allocated to subbranches

Table 2: Summary Statistics

		Equivalized Disposable Household Income		
Country	N	μ	σ	Gini
AT	6,220	25,451	13,971	0.268
BE	6,011	23,291	10,948	0.249
BG	7,154	3,714	2,491	0.333
CH	7,583	42,208	24,486	0.279
CY	4,589	21,058	11,454	0.279
CZ	8,711	9,006	4,320	0.250
DE	12,683	22,221	12,273	0.276
DK	5,897	32,027	13,836	0.232
EE	5,338	6,922	3,912	0.330
EL	6,184	13,184	8,651	0.334
ES	15,481	17,088	10,597	0.329
FI	9,743	27,517	13,891	0.246
FR	11,078	24,299	14,583	0.288
HR	6,969	6,627	3,819	0.306
HU	13,330	5,327	2,863	0.276
IE	4,318	24,867	14,307	0.296
IS	3,684	22,190	9,232	0.210
IT	21,070	18,786	11,730	0.309
LT	5,403	4,774	3,150	0.344
LU	6,765	37,911	19,977	0.271
LV	6,423	5,334	3,618	0.363
MT	4,701	13,006	6,747	0.277
NL	11,411	25,210	11,414	0.235
NO	5,026	43,260	16,971	0.202
PL	15,545	6,103	3,690	0.316
PT	5,899	10,781	7,296	0.334
RO	7,867	2,562	1,646	0.337
SE	6,599	26,346	10,700	0.215
SI	13,183	13,772	5,994	0.225
SK	6,779	7,304	3,416	0.257
UK	7,391	25,936	16,815	0.320

Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: N indicates the total number of observations in the respective country sample. The last three columns refer to the country-specific distribution of equivalized disposable household incomes measured in €. μ indicates the mean, σ the standard deviation, and the last column shows inequality as measured by the Gini coefficient.

based on $\tilde{\omega}^+$. As a consequence, there are differences in the actual sample sizes that are available for the different methods. When comparing inequality of opportunity estimates across methods, we tolerate these differences in sample sizes since we want to compare inequality of opportunity estimates by respecting all methods to the greatest extent. To the contrary, when comparing the out-of-sample performance we use the smallest sample size across methods for all calculations, such that the relative out-of-sample performance cannot be driven by sample size differences or non-random attrition through listwise deletion. A thorough discussion of the sensitivity of all methods to different sample sizes is provided in Appendix A.IV.

Benchmark Methods

We compare our estimates from trees and forests against three benchmark methods that have been proposed in the extant literature.

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 equivalent household income on the following circumstances: father's occupation (10 categories), father's and mother's education (5 categories), area of birth (3 categories), and tenancy status of the household (2 categories). The model specification therefore includes 20 binary variables and resembles the specification used in Palomino et al. (2019).

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. In this application we construct 40 such types. Individuals in type t_m are homogeneous with respect to the educational achievement of their highest educated parent (5 categories) as well as their migration status (2 categories). The latter is indicated by a binary variable for whether the respondent is a first or second generation immigrant. Furthermore, they have fathers working in the same occupation (4 categories). To minimize the frequency of sparsely populated types we divert from the occupational list given in Table 1 by 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. This partition is similar but more parsimonious than the one used in Checchi et al. (2016) who base their analysis on a total of 96 types. Notably, in contrast to Checchi et al. (2016) we exclude age from the list of circumstances since it is fairly controversial whether age qualifies as a circumstance characteristic in the relevant sense.

Lastly, we compare our estimates against the latent class approach as proposed by Li Donni et al. (2015). The eligible set of circumstances is the full set of observable circumstances. For the latent class analysis, we follow Li Donni et al. (2015) and select the number of latent types by minimizing BIC.

Model Performance

In order to assess the prediction accuracy of different models, 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 country in our sample, $N^{-H} = \frac{2}{3}N$ while $N^{H} = \frac{1}{3}N$. We fit our models on the training set and compare their performance on the test set according to the following procedure:

1. Run the chosen models on the training data (for the specific estimation procedures, see section II for trees and forests, and section III for our benchmark methods).

2. Store the prediction functions $\hat{f}^{-H}()$.

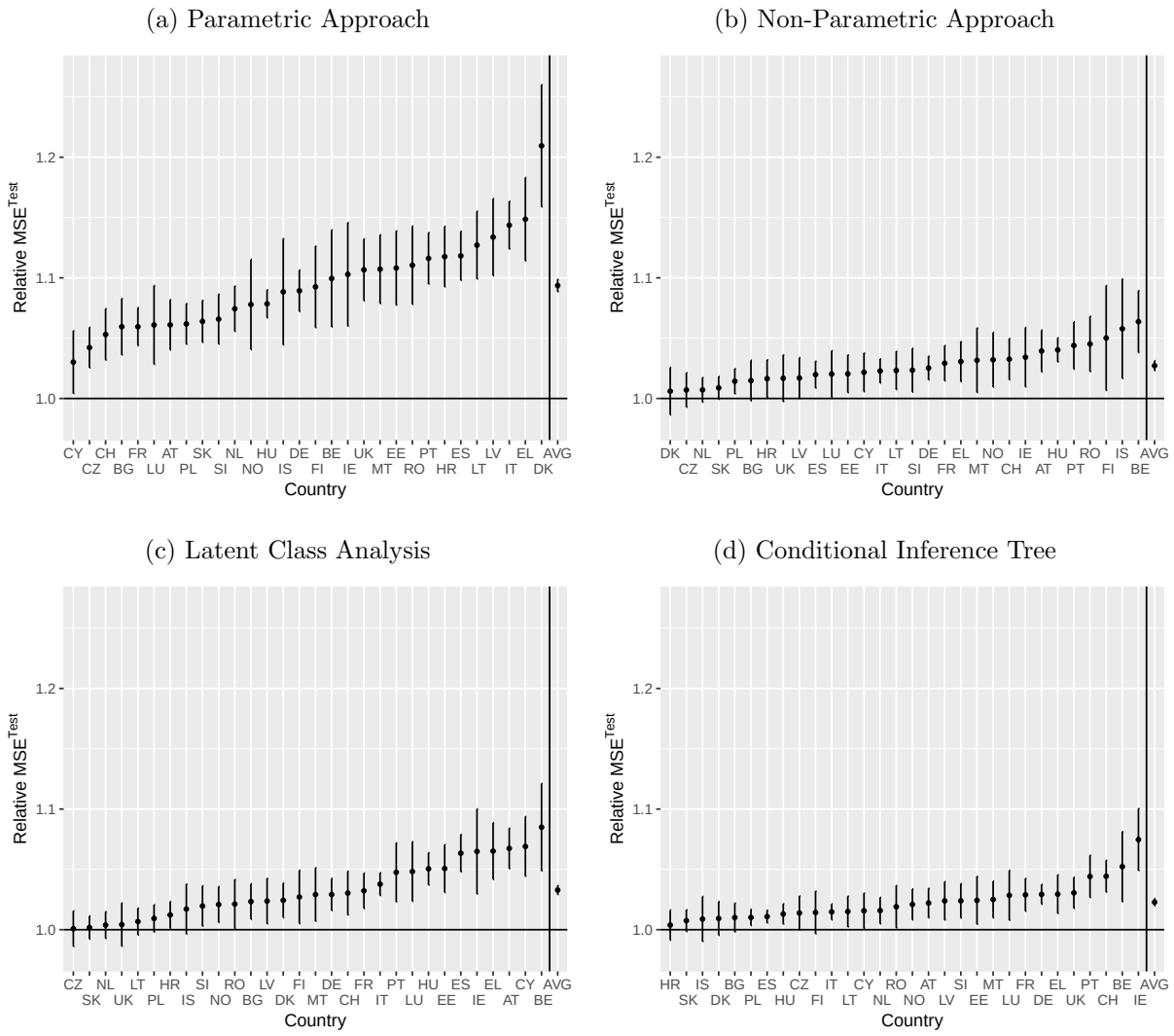
3. Calculate the mean squared error in the test sample:
   $\text{MSE}^{\text{Test}} = \frac{1}{N^{H}} \sum_{i \in H} [y_i - \hat{f}^{-H}(\omega_i)]^2$.

Figure 2 compares the resulting MSE^{Test} of the different models. For each country, MSE^{Test} of random forests is standardized to equal 1, such that an MSE^{Test} larger than 1 represents a worse out-of-sample fit. This implies that the respective method performs worse than forests in trading off upward and downward bias – either by making poor use of circumstance information or overfitting the data. We derive 95% confidence intervals based on 200 bootstrapped re-samples of the test data using the normal approximation method (DiCiccio and Efron, 1996).

Random forests outperform all other methods in all cases. On average, the parametric approach gives a fit that is 9.4% worse than forests (Figure 2, Panel (a)). This average, however, masks considerable heterogeneity. While the relative test error for Cyprus only slightly exceeds the 3%-mark, the test error of the parametric model for Denmark and Sweden exceed the benchmark method by more than 20%. For all countries, the benchmark MSE lies outside the 95% confidence band of the parametric approach.

With average shortfalls of around 3%, out-of-sample prediction errors are less pronounced for the non-parametric (Figure 2, Panel (b)) and latent class models (Figure 2, Panel (c)). Yet, as in the case of the parametric approach, MSE^{Test} statistics of conditional inference forests lie outside the 95% confidence band of the respective method for

Figure 2: Comparison of Models' Test Error



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The y-axis shows the MSE^{Test} of the different estimation approaches relative to the benchmark of random forests. MSE^{Test} for random forests is standardized to 1, such that a relative test error > 1 indicates worse fit than random forests. 95% confidence intervals are derived based on 200 bootstrapped re-samples of the test data using the normal approximation method. For better result visibility Sweden is excluded from the figure since it is an outlier. The test errors for Sweden are 1.43 $[1.21, 1.66]$ for the parametric approach, 1.11 $[1.01, 1.21]$ for the non-parametric approach, 1.06 $[1.02, 1.11]$ for latent class analysis, and 1.06 $[1.01, 1.11]$ for conditional inference trees.

the vast majority of the country cases in our sample. Hence, relative to random forests, the benchmark methods either underutilize or overutilize the information contained in $\mathbf{\Omega}$. As we will see in section III, the parametric and the non-parametric models are overfitting the data and are therefore upward biased. To the contrary, the type partition delivered by latent class analysis tends to be too coarse and therefore downward biased. The relatively good performance of the non-parametric approach could suggest that it is a sustainable alternative to forests. However, since the model specification remains under the discretion of the researcher, this performance is a luck of the draw rather than a property inherent to the estimation approach. In this particular case, had we followed

the specification of Checchi et al. (2016) exactly by incorporating age as a circumstance characteristic, the type partition would more than double and be accompanied with a significant deterioration in the out-of-sample performance (see section III).

On average, conditional inference trees are closest to the test error rate of forests. With the exception of two country cases, the test error of trees exceeds the test error of forests by less than 5%. Yet, as outlined in section II, they also fall short of the performance of forests due to their poorer utilization of the information given in $\mathbf{\Omega}$.

We conclude that among all considered methods, conditional inference forests deliver the highest out-of-sample prediction accuracy. Hence, they perform best in trading off upward and downward bias in inequality of opportunity estimations. One may suspect that other machine learning algorithms perform even better in predicting outcomes out-of-sample. However, we note that in social science applications the gain in prediction accuracy is typically small when alternating between algorithms that allow for sufficient model flexibility. For example, in the context of intergenerational mobility estimations Blundell and Risa (2019) show that there is no difference in the performance of random forests, neural nets and gradient boosted trees.[8] To demonstrate the substantive relevance of this property, we now turn to a comparison of the equality of opportunity estimates emanating from the considered set of estimation approaches.

Estimates of Inequality of Opportunity

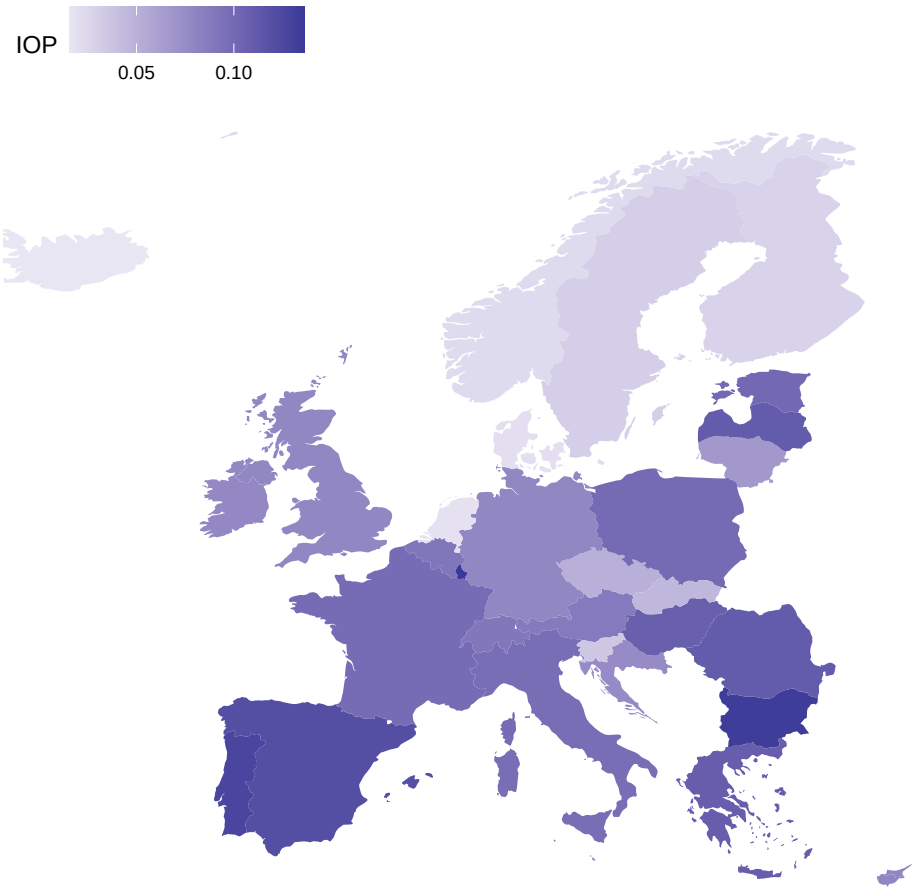
Figure 3 maps inequality of opportunity for our European country sample in 2010. Inequality of opportunity estimates are obtained by calculating the Gini index in the estimated counterfactual distribution $\hat{y}^C$, where the latter is constructed from the predictions of conditional inference forests.[9] We observe a clear North-South gradient with the Scandinavian countries being characterized by the lowest level of inequality of opportunity. Similarly, we observe a slight East-West gradient with many countries from the former

[8] Although it is not explicitly part of our methodological comparison, we provide the exact time necessary to run a single iteration for all countries for each method in the following. (i) Non-parametric approach: 2.45 seconds, (ii) Parametric approach: 1.55 seconds, (iii) Latent class analysis: 1.02 hours, (iv) conditional inference trees: 39.14 seconds, (v) conditional inference forests: 2.06 hours. The run times are measured for a computer with a of 2.3 GHz Intel Core i5 central processor.

[9] As discussed in section I, there is a class of functionals that can be used to summarize the distribution of $\hat{y}^C$. We therefore provide estimates for alternative inequality indexes in Supplementary Material B.

Warsaw pact being characterized by higher levels of inequality of opportunity. Notable exceptions are Czech Republic and Slovakia.

Figure 3: Inequality of Opportunity in Europe, 2010



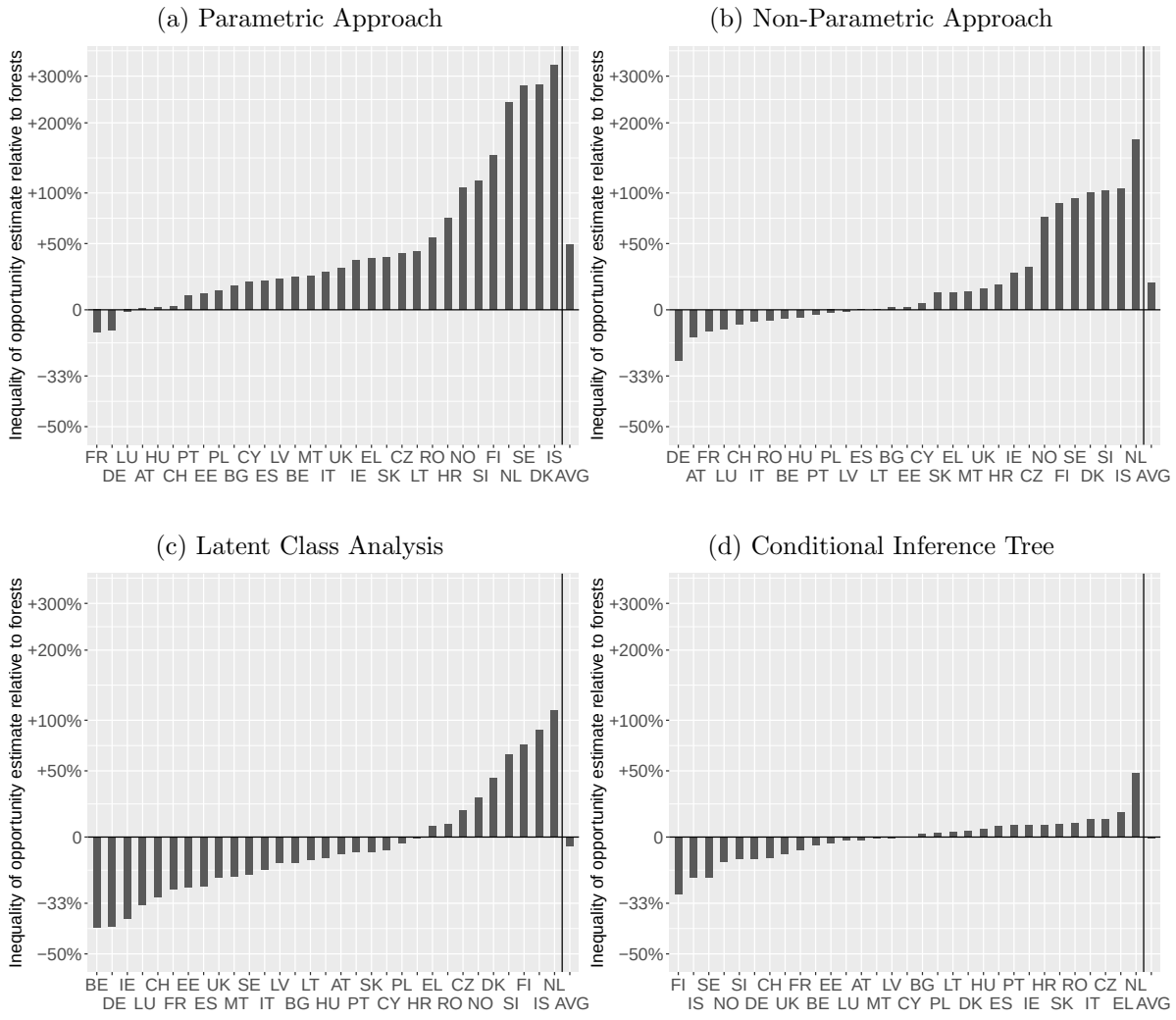
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: Inequality of opportunity is measured by the Gini coefficient in the estimated counterfactual distribution $\hat{y}^C$. $\hat{y}^C$ is constructed based on the predictions from conditional inference forests. Darker shaded colors indicate higher levels of inequality of opportunity. The displayed inequality of opportunity estimates are reported in the last column of Table A.2.

It is important to emphasize that the results of the random forests cannot be interpreted as recovering the truth. However, they provide a benchmark estimate since forests have the lowest test error for all countries, therefore perform best in balancing concerns about upward and downward bias, and hence provide the best *approximation of the truth* among all methods we consider. Following this insight, Figure 4 plots inequality of opportunity estimates based on each method relative to the estimates from conditional inference forests on a logarithmic scale.[10] For each country- and method-specific estimate

[10]Note that the ranking of countries in terms test error rates shown in Figure 2 does not necessarily correspond to the ranking of countries in in terms of the equality opportunity estimates shown in Figure 4. This fact is a consequence of restricting the sample to the smallest available data sample across methods for the assessment of the model performance. See the last paragraph of section III for a detailed explanation.

we divide by the estimate from random forests to obtain the relative divergence between the respective benchmark and our preferred method. This implies that, for a given country, inequality of opportunity estimates larger than those obtained from forests overfit the data and vice versa. An overview table of the underlying point estimates including 95% confidence bands is disclosed in Appendix A.V.

Figure 4: Comparison of Estimates by Method



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: In each panel, the y-axis shows the inequality of opportunity estimate from the method in question divided by the inequality of opportunity estimate from forests, displayed on a logarithmic scale. Country-estimates above the black line indicate an overestimation of inequality of opportunity relative to the random forest benchmark. Reversely, country-estimates below the black line indicate an underestimation of inequality of opportunity relative to the random forest benchmark. For all methods inequality of opportunity is measured by the Gini coefficient in the estimated counterfactual distribution $\hat{y}^C$.

Panel (a) plots the estimates from the parametric approach relative to the forest estimates. For 28 out of 31 countries the inequality of opportunity estimates are higher than the results from conditional inference forests. The given specification of the parametric approach inflates inequality of opportunity statistics by 47% on average. The most pro-

nounced overstatement is observed for Iceland where the parametric approach yields an estimate more than four times higher than the forest analogue. Similarly, the figures of Sweden and Denmark are inflated by a factor of 3.8. Also in terms of country rankings, the parametric approach delivers markedly different results in comparison to our preferred method. While the parametric approach identifies Romania (RO), Bulgaria (BG) and Greece (EL) as the countries in which opportunities are most unequally distributed, these countries rank 6th, 2nd and 7th in the case of forests.

Panel (b) illustrates that the benchmark specification of the non-parametric approach takes a middle-ground between the parametric approach and our preferred method. For 19 out of 31 countries the non-parametric estimate exceeds its forest-based analogue. The non-parametric specification inflates inequality of opportunity statistics at a rate of 18% on average. Also in terms of country rankings the non-parametric approach shows a much closer resemblance to our preferred method than the parametric approach. For example, it identifies Bulgaria (BG), Portugal (PT) and Luxembourg (LU) as the countries in which opportunities are most unequally distributed. This ranking is congruent with the top three countries identified by forests. However, the resemblance should be interpreted as a luck of the draw rather than a property inherent to the estimation approach. Under alternative type partitions the estimates from the non-parametric approach may diverge much more strongly than under the partition adopted in this work.

As shown in Panel (c), the latent class model tends to provide lower estimates than the previous methods. For 22 out of 31 countries the latent class estimate falls short of the forest-based estimate. Given the set of observed circumstances latent class analysis understates inequality of opportunity by 6% on average. The most pronounced understatement of inequality of opportunity is observed for Belgium and Germany. For these countries the latent class model provides estimates more than 40% lower than the forest-based analogues. However, in spite of the tendency to underestimate, there remain four countries for which latent class analysis overstates inequality of opportunity by more than 50% relative to the forest benchmark. Also in terms of country rankings the latent class approach differs markedly from our preferred method. It identifies Romania (RO), Greece (EL) and Portugal (PT) as the countries in which opportunities are most unequally dis-

tributed, whereas these countries rank 6th, 7th and 3rd in the case of forests.

Finally, Panel (d) shows that trees and forests tend to produce similar results. The correlation between estimates is high (0.98) and in contrast to all other approaches there is no general tendency to over- or underestimate inequality of opportunity relative to random forests. In view of the discussed shortcomings of trees, it is unsurprising that some estimates divert from their forest-based analogues. However, even the most notable outliers – Finland at the lower end, and the Netherlands at the upper end – remain well below the extrema of the benchmark methods considered previously.

To summarize: according to our benchmark specifications the parametric and the non-parametric approach tend to overestimate inequality of opportunity. To the contrary, estimates based on latent class analysis tends to underestimate inequality of opportunity. The poor out-of-sample replicability of standard estimation approaches in conjunction with the large divergences of their inequality of opportunity estimates from approaches that perform better in the first dimension, illustrate the importance of appropriate model specifications when comparing societies with respect to their need for opportunity equalizing policy interventions.

Opportunity Structure

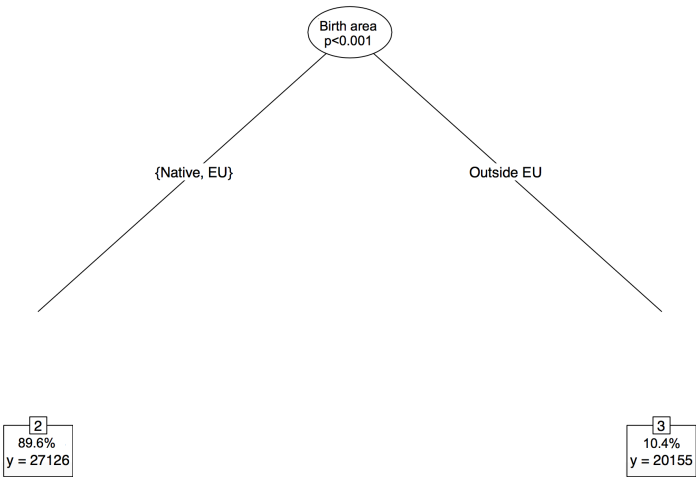
Endowed with an estimate of inequality of opportunity, adequate policy responses must be informed by the opportunity structure of a society. Policymakers want to learn about the particular circumstance characteristics which drive the existence of inequality of opportunity. In this section we illustrate such analyses for both trees and forests. To keep the analysis intelligible we restrict ourselves to two interesting country cases: Sweden and Germany. Readers interested in the opportunity structures of the remaining 29 countries are referred to Supplementary Material C.

We are careful to emphasize that one cannot ascribe any causality to our estimates. However, in spite of the correlative nature of the displayed opportunity structures, they may provide useful starting points for decisionmakers to locate policy areas for opportunity equalizing reforms or to stimulate further academic investigation by means of detailed decomposition or causal analyses (Fortin et al., 2011). In the case of trees, it is also worth-

while to keep in mind that their structure remains rather sensitive to small perturbations of the data. In this application, however, tree structures are affirmed by variable importance calculations based on forests which are less sensitive to such perturbations. This validation is a tentative confirmation that graphical tree representations can serve as useful starting points for the analysis of opportunity structures.

Trees. Figure 5 illustrates that the opportunity structure of Sweden can be summarized by a tree with two terminal nodes. Inequality of opportunity in Sweden is due to marked differences between first-generation immigrants born outside of Europe and the collective group of native residents and European immigrants. The former group accounts for about 10% of the population and on average obtains an equivalent household income that is 26% lower than the corresponding income of the latter group. The between-type Gini is 0.025 or about 12% of total inequality. We note that our estimates differ from Björklund et al.

Figure 5: Opportunity Tree (Sweden)



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ 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 expecatation y^C.

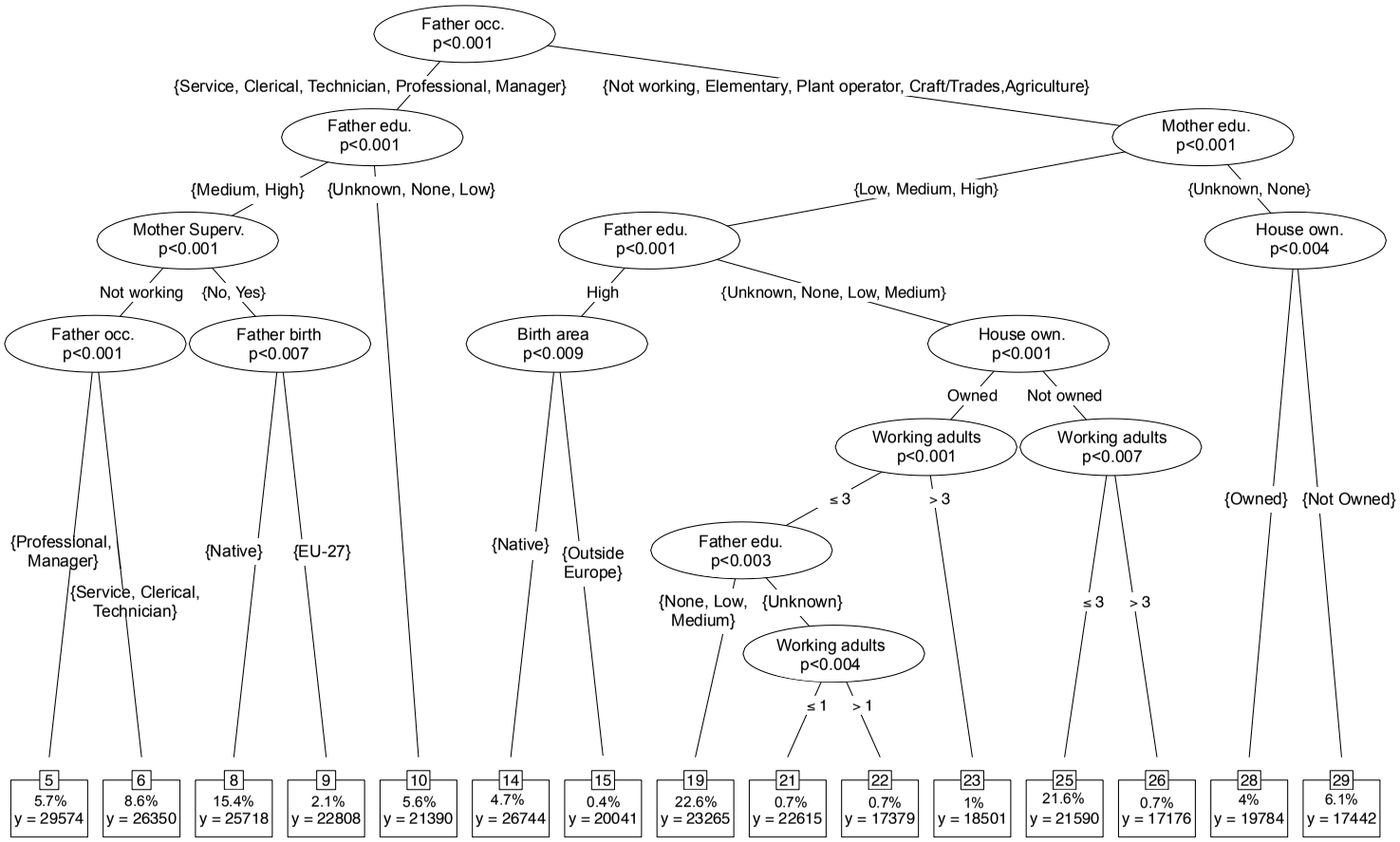
(2012a) who use Swedish registry data to estimate inequality of opportunity at about 28% of total inequality. These estimates, however, are not strictly comparable to ours since Björklund et al. (2012a) focus on a younger (32-38) male-only sample and market income instead of disposable household income.

A different picture arises when considering Germany. Parental occupation, parental

education, migration status, the number of working adults in the household, and parental tenancy status interact in creating a complex tree made of 14 splits and 15 terminal nodes. The null hypothesis of equality of opportunity is most firmly rejected for individuals whose fathers work in different occupations. If a respondent's father worked in one of the higher ranked occupations (I-01–I-05), the individual belongs to a more advantaged circumstance type than otherwise (Terminal nodes 5-10). These types together account for 37.4% of the population and have an average outcome of €26,380 – far above the population average of €22,221. However, the advantage of this circumstance characteristic is contingent on the educational status of the father. If a respondent's father had no or low education, the offspring earned less (€21,390) than the country average in spite of the fact that fathers made a career in a high-rank occupation. Conditional on the father both being highly educated and working in a high-rank occupation, the intra-household division of labor plays an important role. On the one hand, those individuals coming from single-earner households in which the mother stayed at home are the most advantaged circumstance types of Germany in 2010 – especially so if their father worked as a manager or professional (Terminal nodes 5 and 6). On the other hand, offspring of double-earner households tend to be differentiated by their migration status. Comparing terminal nodes 8 and 9 we learn that the advantage of coming from a highly-educated double-earner household is substantially diminished from €25,718 to €22,808 if the respondent's father was born outside of Germany. A similar distinction based on migration status can be observed on the right-hand side of the tree, in which individuals were born to fathers with a lower occupational status (I-05–I-00). Individuals in this group lived in above average income households if both of their parents were fairly educated *and* their father had no migration background (Terminal node 14). This advantage again vanishes substantially if the respondent's father was born outside of Europe (Terminal node 15).

There is marked heterogeneity in tree structures across countries (Supplementary Material C). For the remaining countries in our sample, terminal nodes range from three (Denmark, Iceland and Norway) to 27 (Italy). It is noteworthy that the rank-rank correlation between the number of terminal nodes and the inequality of opportunity estimates presented in section III is positive but not perfect. Whether a split is conducted is a func-

Figure 6: Opportunity Tree (Germany)



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ 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 type mean ($\hat{\mu}_m$).

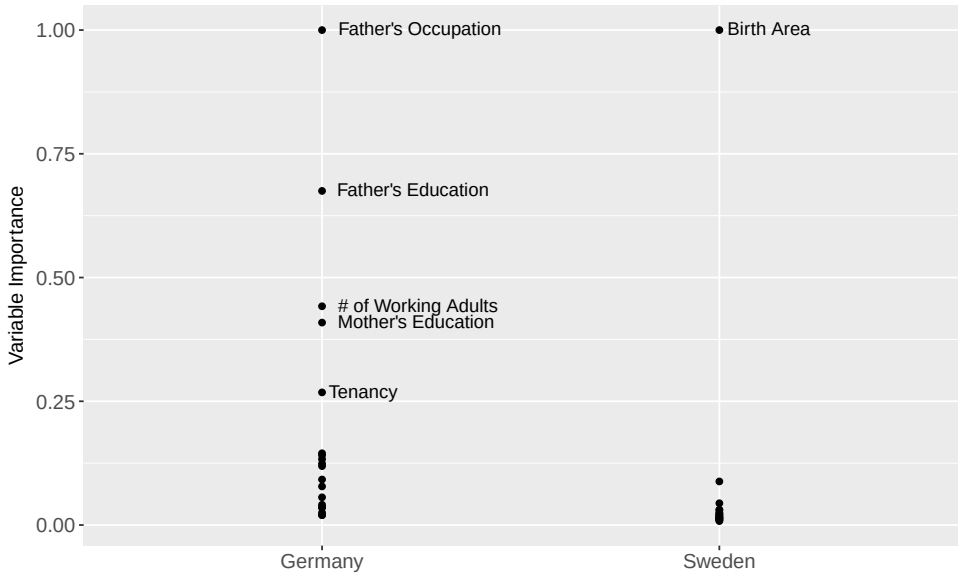
tion of the average income difference and the sample size of the ensuing types. Hence, if the sample size is large enough, the statistical tests underlying the splitting algorithm have sufficient power to detect even minor differences in average incomes across groups. Such small differences, however, have little impact on inequality in the estimated counterfactual distribution $\hat{y}^C$.

Forests. Forests cannot be analyzed in the straightforward graphical manner of trees. However, we can use variable importance measures to assess the impact of circumstance variables for the construction of opportunity forests. One measure of variable importance, as proposed by Strobl et al. (2007), is obtained by permuting input variable ω^p such that its dependence with y is lost. After this, the out-of-bag error rate MSE^{OOB} is re-computed. The increase of MSE^{OOB} in comparison to the baseline out-of-bag error indicates the importance of the input variable for prediction accuracy. Repeating this procedure for all $\omega^p \in \mathbf{\Omega}$ affords a relative comparison of the importance of all circumstances.

Figure 7 shows the results from this procedure for our example cases of Germany and Sweden. Each black dot is the importance of one of the variables in the set of observed circumstances $\mathbf{\Omega}$. We standardize the ensuing results such that the variable importance measure for the circumstance with the greatest impact in each country equals one. For the case of Sweden birth area is the only circumstance that has a meaningful predictive value. In Germany, father's occupation and father's education are most important, followed by the number of working adults in the household and mother's education.

It is reassuring that these findings are in line with the graphical analysis of opportunity trees. In Supplementary Material C we show variable importance plots for all countries in our sample as well maps that group countries by their most important circumstance variable. Broadly, we can divide our country sample into three groups according to the circumstances that determine their opportunity structure. First, there is a handful of primarily Nordic countries where the respondent's birth area is the most important circumstance. Second, there is a large group of primarily Western and Southern European countries where father's occupation and father's education are most important. Third, there is a group of Eastern European countries where mother's education and occupation are most important.

Figure 7: Variable Importance for Germany and Sweden



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. The forests are constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1.

These results have important implications for analyses that use inequality of opportunity estimates as left-hand side variables. Researchers have become increasingly interested in the opportunity equalizing properties of specific policy reforms (e.g. Andreoli et al., 2019). Our results suggest that a one-size-fits-all approach is insufficient to capture the underlying opportunity structures in different societies. Hence, one should be cautious in comparing equality of opportunity estimates based on the same model within a particular country before and after a change in institutional configurations due to a policy reform. While the coefficients on particular circumstance characteristics may change in the course of a reform, the *relevant* model $f()$ may change as well. Therefore, to the extent that researchers are interested in the aggregate opportunity-equalizing effect of a particular reform, they need to take both of these channels into account.

IV 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 performance. Hence, they deliver the best estimates of inequality of opportunity. Conditional inference trees, on the other hand, are econometrically less complex and provide a handy graphical illustration that can be used to analyze opportunity structures. The fact that trees are very close to forests in terms of their out-of-sample performance, their inequality of opportunity estimates, and the importance they assign to specific circumstances makes us confident that they are a useful tool for communicating issues related to inequality of opportunity to a larger audience.

To be sure, the development of machine learning algorithms and their integration into the analytical toolkit of economists is a highly dynamic process. We are well aware that finding the best machine learning algorithm for inequality of opportunity estimations is a methodological horse race with frequent entry of new competitors 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 (Friedberg et al., 2018) that outperform more traditional forest algorithms in their ability to capture the linear impact of particular predictor variables.

Finally, while we restricted ourselves to ex-ante utilitarian measures of inequality of opportunity, the exploration of these algorithms for other methods in the inequality of opportunity literature, such as ex-post measures à la Pistolesi (2009) or ex-ante and ex-post tests à la Kanbur and Snell (2019) and Lefranc et al. (2009), provides another interesting avenue for future research.

References

Aaberge, R., M. Mogstad, and V. Peragine (2011). "Measuring long-term inequality of opportunity". *Journal of Public Economics* 95 (3–4), pp. 193–204.

Alesina, A., S. Stantcheva, and E. Teso (2018). "Intergenerational Mobility and Preferences for Redistribution". *The American Economic Review* 108 (2), pp. 521–554.

Almås, I., A. W. Cappelen, J. T. Lind, E. Ø. Sørensen, and B. Tungodden (2011). "Measuring unfair (in)equality". *Journal of Public Economics* 95 (7–8), pp. 488–499.

Andreoli, F., T. Havnes, and A. Lefranc (2019). "Robust inequality of opportunity comparisons: Theory and application to early-childhood policy evaluation". *The Review of Economics and Statistics* 101 (2), pp. 355–369.

Arneson, R. (2018). "Four conceptions of equal opportunity". *The Economic Journal* 128 (612), F152–F173.

Athey, S. (2018). "The Impact of Machine Learning on Economics". In: *The Economics of Artificial Intelligence: An Agenda*. Ed. by A. K. Agrawal, J. Gans, and A. Goldfarb. Chicago: University of Chicago Press. Chap. 21.

Athey, S. and G. Imbens (2019). "Machine Learning Methods Economists Should Know About". *Annual Review of Economics* 11, pp. 685–725.

Belhaj Hassine, N. (2011). "Inequality of Opportunity in Egypt". *The World Bank Economic Review* 26 (2), pp. 265–295.

Biau, G. and E. Scornet (2016). "A random forest guided tour". *Test* 25 (2), pp. 197–227.

Björklund, A., M. Jäntti, and J. E. Roemer (2012a). "Equality of opportunity and the distribution of long-run income in Sweden". *Social Choice and Welfare* 39 (2-3), pp. 675–696.

Björklund, A., J. Roine, and D. Waldenström (2012b). "Intergenerational top income mobility in Sweden: Capitalist dynasties in the land of equal opportunity?" *Journal of Public Economics* 96 (5-6), pp. 474–484.

Blau, F. D. and L. M. Kahn (2017). "The Gender Wage Gap: Extent, Trends, and Explanations". *Journal of Economic Literature* 55 (3), pp. 789–865.

Blundell, J. and E. Risa (2019). "Income and family background: Are we using the right models?" *mimeo*.

Bourguignon, F., F. H. G. Ferreira, and M. Menéndez (2007). "Inequality of Opportunity in Brazil". *Review of Income and Wealth* 53 (4), pp. 585–618.

Braun, S. T. and J. Stuhler (2018). "The Transmission of Inequality Across Multiple Generations: Testing Recent Theories with Evidence from Germany". *The Economic Journal* 128 (609), pp. 576–611.

Breiman, L., J. Friedman, C. Stone, and R. Olshen (1984). *Classification and Regression Trees*. Belmont: Taylor & Francis.

Breiman, L. (2001). "Random forests". *Machine Learning* 45 (1), pp. 5–32.

Brunori, P., V. Peragine, and L. Serlenga (2019). "Upward and downward bias when measuring inequality of opportunity". *Social Choice and Welfare* 52 (4), pp. 635–661.

Cappelen, A. W., A. D. Hole, E. Ø. Sørensen, and B. Tungodden (2007). "The Pluralism of Fairness Ideals: An Experimental Approach". *The American Economic Review* 97 (3), pp. 818–827.

Checchi, D. and V. Peragine (2010). "Inequality of opportunity in Italy". *The Journal of Economic Inequality* 8 (4), pp. 429–450.

Checchi, D., V. Peragine, and L. Serlenga (2016). "Inequality of Opportunity in Europe: Is There a Role for Institutions?" In: *Inequality: Causes and Consequences*. Ed. by L. Cappellari, S. Polachek, and K. Tatsiramos. Vol. 43. Research in Labor Economics. Bingley: Emerald, pp. 1–44.

Chetty, R., N. Hendren, and L. F. Katz (2016). "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment". *The American Economic Review* 106 (4), pp. 855–902.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014a). "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States". *The Quarterly Journal of Economics* 129 (4), pp. 1553–1623.

Chetty, R., N. Hendren, P. Kline, E. Saez, and N. Turner (2014b). "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility". *The American Economic Review* 104 (5), pp. 141–47.

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

Corak, M. and P. Piraino (2011). "The Intergenerational Transmission of Employers". *Journal of Labor Economics* 29 (1), pp. 37–68.

Cowell, F. (2016). "Inequality and Poverty Measures". In: *Oxford Handbook of Well-Being and Public Policy*. Ed. by M. D. Adler and M. Fleurbaey. Oxford: Oxford University Press. Chap. 4, pp. 82–125.

Cowell, F. and M.-P. Victoria-Feser (1996). "Robustness Properties of Inequality Measures". *Econometrica* 64 (1), pp. 77–101.

DiCiccio, T. J. and B. Efron (1996). "Bootstrap Confidence Intervals". *Statistical Science* 11 (3), pp. 189–212.

Felfe, C. and R. Lalive (2018). "Does early child care affect children's development?" *Journal of Public Economics* 159, pp. 33–53.

Ferreira, F. H. G. and J. 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.

Ferreira, F. H. G., C. Lakner, M. A. Lugo, and B. Özler (2018). "Inequality of Opportunity and Economic Growth: A Cross-Country Analysis". *Review of Income and Wealth* 64 (4), pp. 800–827.

Fleurbaey, M. (1995). "Three solutions for the compensation problem". *Journal of Economic Theory* 65 (2), pp. 505–521.

Fortin, N., T. Lemieux, and S. Firpo (2011). "Decomposition Methods in Economics". In: *Handbook of Labor Economics*. Ed. by D. Card and O. Ashenfelter. Vol. Volume 4, Part A. Amsterdam: Elsevier. Chap. 1, pp. 1–102.

Friedberg, R., J. Tibshirani, S. Athey, and S. Wager (2018). "Local Linear Forests". *arXiv* 1807.11408.

Friedman, J., T. Hastie, and R. Tibshirani (2009). *The elements of statistical learning*. New York: Springer.

García, J. L., J. J. Heckman, and A. L. Ziff (2018). "Gender Differences in the Benefits of an Influential Early Childhood Program". *European Economic Review* 109 (October 2018), pp. 9–22.

Hothorn, T., K. Hornik, and A. Zeileis (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework". *Journal of Computational and Graphical Statistics* 15 (3), pp. 651–674.

Hufe, P., A. Peichl, J. E. Roemer, and M. Ungerer (2017). "Inequality of Income Acquisition: The Role of Childhood Circumstances". *Social Choice and Welfare* 143 (3–4), pp. 499–544.

Jusot, F., S. Tubeuf, and A. Trannoy (2013). "Circumstances and Efforts: How Important is their Correlation for the Measurement of Inequality of Opportunity in Health?" *Health Economics* 22 (12), pp. 1470–1495.

Kanbur, R. and A. Snell (2019). "Inequality Measures as Tests of Fairness". *The Economic Journal* 129 (621), pp. 2216–2239.

Kreisman, D. and M. A. Rangel (2015). "On the Blurring of the Color Line: Wages and Employment for Black Males of Different Skin Tones". *The Review of Economics and Statistics* 97 (1), pp. 1–13.

Lefranc, A., N. Pistolesi, and A. 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, P., J. G. Rodríguez, and P. 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.

Mare, R. D. (2011). "A Multigenerational View of Inequality". *Demography* 48 (1), pp. 1–23.

Marrero, G. A., J. G. Rodríguez, and R. Van Der Weide (2016). "Unequal opportunity, unequal growth". *World Bank Policy Research Working Paper* 7853.

Morgan, J. N. and J. 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, S. and J. Spiess (2017). "Machine Learning: An Applied Econometric Approach". *Journal of Economic Perspectives* 31 (2), pp. 87–106.

Palomino, J. C., G. A. Marrero, and J. G. Rodríguez (2019). "Channels of inequality of opportunity: The role of education and occupation in Europe". *Social Indicators Research* 143, pp. 1045–1074.

Peragine, V., F. Palmisano, and P. Brunori (2013). "Economic Growth and Equality of Opportunity". *The World Bank Economic Review* 28 (2), pp. 247–281.

Pistolesi, N. (2009). "Inequality of opportunity in the land of opportunities, 1968–2001". *The Journal of Economic Inequality* 7 (4), pp. 411–433.

Ramos, X. and D. Van de gaer (2016). "Empirical Approaches to Inequality of Opportunity: Principles, Measures, and Evidence". *Journal of Economic Surveys* 30 (5), pp. 855–883.

Roemer, J. E. (1998). *Equality of Opportunity*. Cambridge: Harvard University Press.

Roemer, J. E. and A. Trannoy (2015). "Equality of Opportunity". In: *Handbook of Income Distribution*. Ed. by A. B. Atkinson and F. Bourguignon. Vol. 2. Amsterdam: Elsevier. Chap. 4, pp. 217–300.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution". *BMC bioinformatics* 8 (1), p. 25.

Van de gaer, D. (1993). "Equality of Opportunity and Investment in Human Capital". PhD thesis. University of Leuven.

Varian, H. R. (2014). "Big Data: New Tricks for Econometrics". *Journal of Economic Perspectives* 28 (2), pp. 3–27.

Vosters, K. (2018). "Is the Simple Law of Mobility Really a Law? Testing Clark's hypothesis". *The Economic Journal* 128 (612), F404–F421.

Vosters, K. and M. Nybom (2018). "Intergenerational Persistence in Latent Socioeconomic Status: Evidence from Sweden and the United States". *Journal of Labor Economics* 35 (3), pp. 869–901.

A APPENDIX

A.I Conditional Inference Algorithm

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 $\omega^p \in \mathbf{\Omega}$, 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).

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

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

   $\Rightarrow$ 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. $\tilde{\omega}^* = \underset{\omega_s^*}{\operatorname{argmin}}\{p^{\omega_s^*} : \omega_s^* \in \mathbf{\Omega}\}$.

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

A.II 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 (Friedman et al., 2009) – provides unbiased estimates of the out-of-sample MSE. To perform cross-validation, one starts by splitting the sample into K roughly equal-sized folds. Then, one implements the conditional inference algorithm on the union of $K - 1$ folds for varying levels of α, while leaving out the kth subsample. This makes it possible to compare the predictions

emanating from the $K-1$ folds with the unused data points observed in the kth fold. One then calculates the out-of-sample MSE as a function of α:

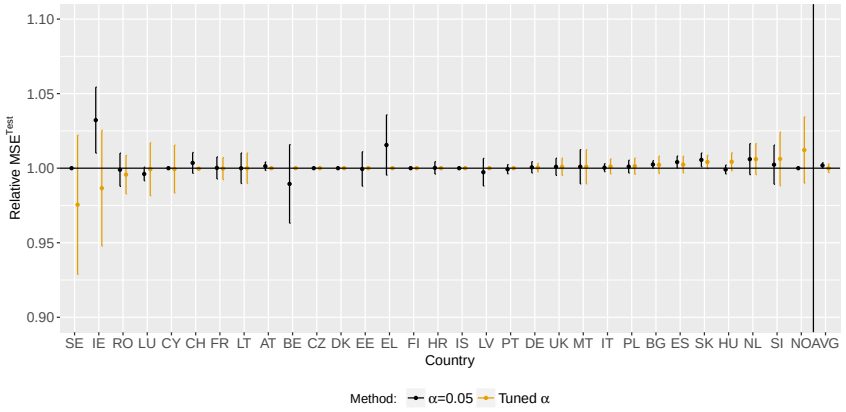
$$\text{MSE}_k^{\text{CV}}(\alpha) = \frac{1}{N^k}\sum_{i \in k}(y_i^k - \hat{f}^{-k}(\omega_i;\alpha))^2,\ \omega_i \in \mathbf{\Omega},\ i \in \mathbf{N}, \tag{9}$$

where $\hat{f}^{-k}()$ denotes the estimation function $\hat{f}()$ constructed while leaving out the kth fold. Note that every fold may render a new $\hat{f}()$. This exercise is repeated for all K folds, so that $\text{MSE}^{\text{CV}}(\alpha) = \frac{1}{K}\sum_k \text{MSE}_k^{\text{CV}}(\alpha)$. One then chooses α^* such that

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

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

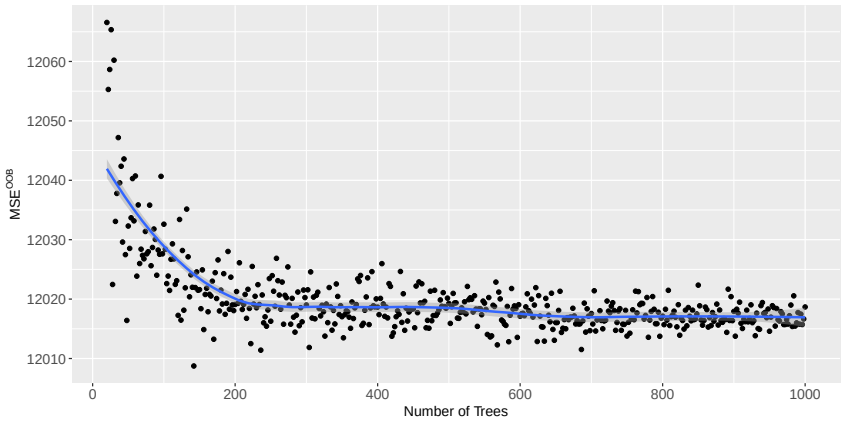
Figure A.1: Tuning Conditional Inference Trees



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The y-axis shows the MSE^{Test} for different specifications of α^* relative to the baseline specification of $\alpha^* = 0.01$. The MSE^{Test} for the baseline specification of $\alpha^* = 0.01$ is standardized to 1, such that a relative test error > 1 indicates worse fit than the baseline specification. 95% confidence intervals are derived based on 200 bootstrapped re-samples of the test data using the normal approximation method. When no confidence intervals are shown, the methods give the same MSE^{Test}. Trees are constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference trees is detailed in Table 1. For the construction of MSE^{Test}, see section III. Black dots and the associated confidence bands show results for $\alpha^* = 0.05$. Orange dots and the associated confidence bands show results for cross-validated α^* using $K = 5$ folds.

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, to reduce computational costs 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

Figure A.2: Optimal Size of Forests



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The x-axis shows the parameter value for B^*, i.e. the number of trees per forest. The dots show the MSE^{OOB} obtained from estimating a random forest with the given number of trees for the case of Germany. We allow for 6 circumstances to be considered at each splitting point ($\bar{P}^* = 6$). Due to the randomness in the observations selected for each tree and the randomness in the circumstances allowed at each splitting point, even when estimating multiple forests with the same number of trees, the associated MSE^{OOB} will differ. The blue line is a non-parametric fitted line of the MSE^{OOB} estimates and the shaded area the 95% confidence interval of this line. Evidently, as the tree size approaches 200, on expectation, the MSE^{OOB} stops improving much.

with $B^* = 200$ for most countries in our sample (Figure A.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 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:
   $\text{MSE}^{\text{OOB}}(\alpha, \bar{P}) = \frac{1}{N} \sum_i [y_i - \hat{f}^{\text{OOB}}(\omega_i; \alpha, \bar{P})]^2$.

One then chooses the combination of parameter values that delivers the lowest MSE^{OOB}:

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

The logic behind this tuning exercise is similar to cross-validation. However, instead of leaving out the kth fraction of the dataset to make out-of-sample predictions, we leverage

the fact that each tree of a forest is grown on a subsample $\mathcal{S}' \subset \mathcal{S}$ that excludes some observations $i \in \{1, ..., S\}$. Hence, for each tree we can use the out-of-bag data points to evaluate the predictive accuracy of the respective model. Using this out-of-bag procedure, the optimal level of α is often very high, meaning that the trees are grown very deep. At the limit, α may be set to 1, in which case splits are made as long as each end node has at least a minimum number of observations. If we were to extract a single tree from such a forest, then this tree would surely overfit the data and perform poorly out-of-sample. However, when averaging over many overfitted trees this drawbacks gets remedied. Hence, setting $\alpha^* = 1$ a priori is a sensible strategy to reduce the computational cost of random forests.

A.III Upward Bias, Downward Bias and the MSE

A standard statistic adopted 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 individual's conditional expectation $\mathbb{E}(y|\omega)$ in a random sample $\mathcal{S}$. The MSE can be decomposed into three components (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 bias-variance decomposition. Components (1) and (2) can be directly linked to concerns of upward and downward biases in inequality of opportunity estimates. To illustrate this point, notice that we minimize (1) by imposing the following model specification $y = \hat{f}(\omega) + \epsilon = \beta_0 + \epsilon$. This model specification assumes that individual outcomes are best predicted by the sample mean $\mu^{\mathcal{S}}$. For the sake of illustration, furthermore assume that each population sample $\mathcal{S}$ is large enough such that its mean corresponds to the underlying population truth: $\mu^{\mathcal{S}} = \mu$. Obviously, this is a stark assumption which we only make for illustration purposes. In reality, there will always be some variance in the sample means as long as one does not capture the entire underlying population. As a consequence, (1) drops out and the MSE is entirely captured by components (2) and (3):

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

This shows that the variance-minimizing model is established by the assumption that everybody in the population faces exactly the same circumstance set, i.e. $\omega_i = 1 \; \forall i \in \mathbf{N}$ – and hence that the value of every individual opportunity set is best estimated by the sample mean $\mu^{\mathcal{S}}$. Under the given assumptions, the model cannot give an upward biased estimate of inequality of opportunity since it is restricted in a way that does not

allow for any role of circumstance characteristics in the explanation of individual outcome differences. In fact, for any functional $I()$ that satisfies the measurement criteria outlined in section I, $I(\hat{y}^C) = 0$.

Reversely, one could ask which model would minimize component (2) of the MSE. To this end, we would have to specify a complex model that allows for the full set of relevant circumstances, their mutual interactions and non-linearities such that *in expectation* we would obtain unbiased estimates of $f(\omega)$. In this case the MSE would be entirely captured by components (1) and (3):

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

While such a model in expectation provides unbiased estimates of y^C, the conditional expectations within a particular sample $\mathcal{S}$ may be estimated with error: $\hat{y}_i^C = y_i^C + u_i$. u_i is an iid error the importance of which tends to increase with model complexity (Friedman et al., 2009). Hence, model complexity leads to measurement error which in turn inflates the variance of $\hat{y}^C$ in comparison to the underlying truth: $\text{Var}(\hat{y}^C) = \text{Var}(y^C) + \text{Var}(u)$. As shown in Brunori et al. (2019), applying any functional $I()$ that satisfies the measurement criteria outlined in section I to the variance inflated distribution $\hat{y}^C$ results in upward biased estimates of inequality of opportunity.

A.IV Sensitivity to Sample Size

Table A.1 shows the country sample sizes from our empirical application (Section III) by estimation method. Note that the parametric and the non-parametric approach tend to have smaller sample sizes since they rely on list-wise deletion of observations in case of item non-response.

Table A.1: Sample Size by Method

Country	Parametric Approach	Non-parametric Approach	CI Forests and Trees
AT	6,042	6,107	6,220
BE	4,528	5,375	6,011
BG	5,952	6,210	7,154
CH	6,420	6,754	7,583
CY	4,483	4,525	4,589
CZ	6,438	6,524	8,711
DE	10,539	11,139	12,683
DK	2,107	2,223	5,897
EE	4,857	5,004	5,338
EL	5,743	5,862	6,184
ES	14,640	14,816	15,481
FI	2,900	3,207	9,743
FR	10,104	10,391	11,078
HR	5,945	6,159	6,969
HU	12,139	12,525	13,330
IE	3,080	3,138	4,318
IS	1,447	1,492	3,684
IT	20,238	20,800	21,070
LT	4,539	4,703	5,403
LU	6,528	6,654	6,765
LV	6,046	6,192	6,423
MT	4,048	4,117	4,701
NL	5,414	5,518	11,411
NO	2,329	2,400	5,026
PL	12,676	13,182	15,545
PT	5,689	5,795	5,899
RO	5,701	6,145	7,867
SE	467	561	6,599
SI	4,691	4,747	13,183
SK	6,170	6,401	6,779
UK	5,756	5,922	7,391
Minimum (N^{min})	467	561	3,684

Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: Each column refers to the number of observations used in the estimation of inequality of opportunity for the particular approach.

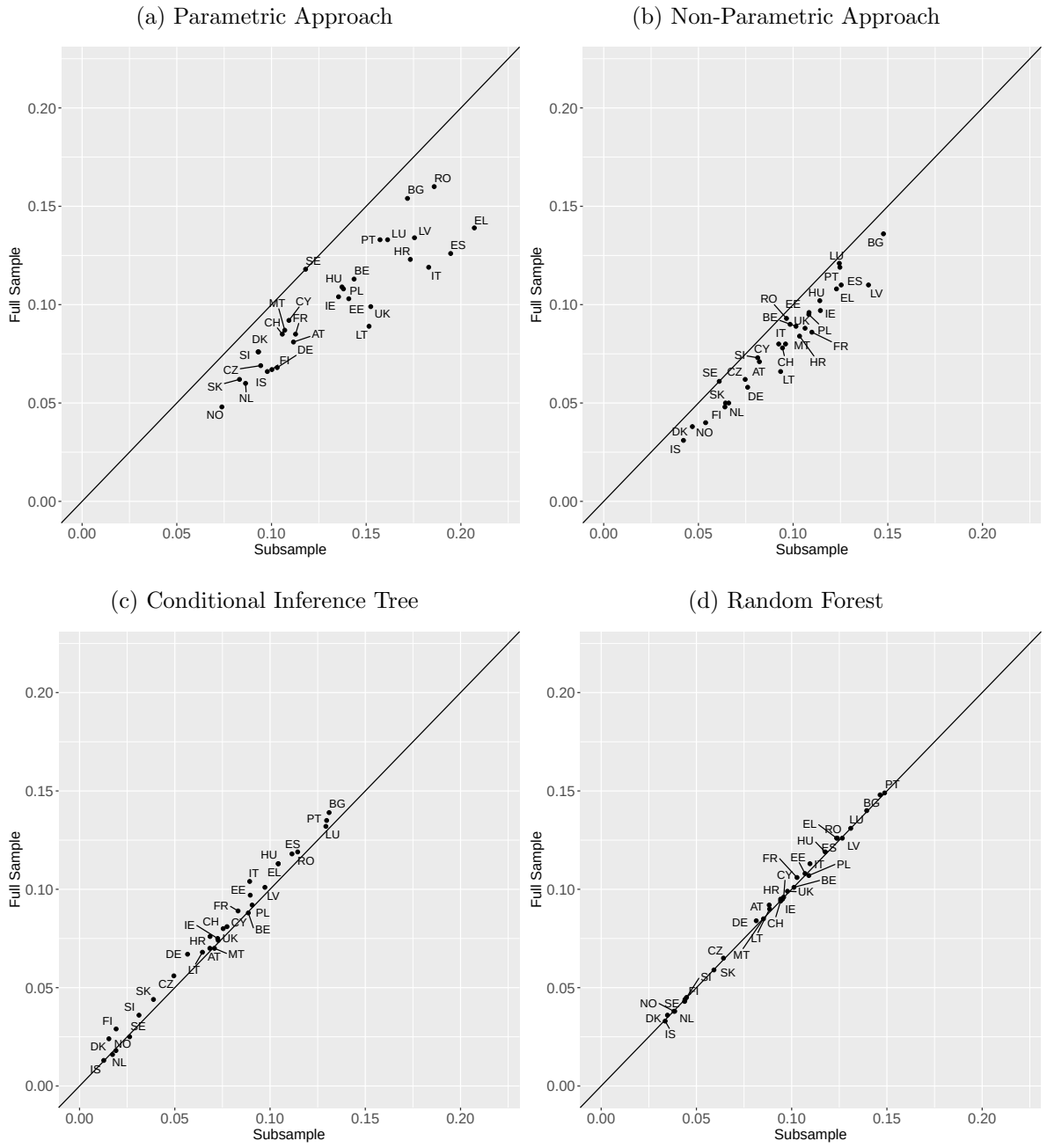
To analyze the extent to which inequality of opportunity estimates are a function of sample sizes we rely on the following procedure.

1. For each country-method cell we make 10 random draws from the full country sample. The size of each subsample is determined by the smallest method-specific country sample: $N^{\text{min}}_{\text{trees}} = N^{\text{min}}_{\text{forests}} = 3,684$, $N^{\text{min}}_{\text{non-parametric}} = 561$ and $N^{\text{min}}_{\text{parametric}} = 467$.

2. We estimate inequality of opportunity on each of the 10 subsamples and average over these 10 iterations to obtain an estimate for each country-method cell.

Figure A.3 plots the estimates based on the full sample against the estimates from the subsamples as derived from the procedure outlined above.

Figure A.3: Inequality of opportunity with full sample and random subsamples



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: In each panel the x-axis indicates the country-specific inequality of opportunity estimate of the respective estimation approach on the subsample $N^{\min}$, which differs by method. Analogously, the y-axis indicates the country-specific inequality of opportunity estimate of the respective estimation approach on the full sample N. For all methods inequality of opportunity is measured by the Gini coefficient in the estimated counterfactual distribution $\hat{y}^C$. The black solid line is the 45-degree line. Country-estimates above the 45-degree line indicate an underestimation of inequality of opportunity relative to the full sample benchmark. Reversely, country-estimates below the 45-degree line indicate an overestimation of inequality of opportunity relative to the full sample benchmark. To avoid the intricacies of tuning forests, we set $\bar{P}^* = 8$ and $\alpha^* = 1$. As a consequence, our benchmark estimates may slightly differ from the ones reported in the main text.

Panels (a) and (b) illustrate that the parametric and non-parametric approaches tend to overestimate inequality of opportunity as sample sizes decline. This is a direct consequence of fixing the number of model parameters a priori. As the available degrees of

freedom decline, the model tends to overfit, which translates into upward biased estimates of inequality of opportunity (see Appendix Section A.III). To the contrary, in the case of trees and forests (Panels (c) and (d)) country estimates align closely along the 45-degree line. This pattern illustrates that trees and forests are less sensitive to variations in the sample size. On the one hand, if the sample size is small, the p-values of the hypothesis tests will increase and less splits will be conducted. This prevents the models from overfitting and safeguards the inequality of opportunity estimate against overestimation. On the other hand, both methods allow for extremely flexible functional forms in the construction of $\hat{y}^C$. Hence, even when reducing the sample size by a factor of 5.8 (Italy) the estimate from the subsample closely aligns with the inequality of opportunity estimate from the full country sample.

We conclude that inequality of opportunity comparisons across countries based on trees and forests are more robust to sample size variations than alternative estimation approaches.

A.V Point Estimates and Confidence Intervals

Table A.2: Inequality of Opportunity Estimates

	Benchmark Methods			Conditional Inference	
Country	Parametric	Non-Parametric	Latent Class	Tree	Random Forest
AT	0.089 [0.081;0.097]	0.075 [0.067;0.083]	0.080 [0.070;0.090]	0.087 [0.076;0.097]	0.088 [0.080;0.096]
BE	0.111 [0.100;0.121]	0.087 [0.080;0.094]	0.053 [0.036;0.071]	0.087 [0.078;0.096]	0.091 [0.084;0.098]
BG	0.154 [0.145;0.163]	0.136 [0.126;0.145]	0.115 [0.106;0.124]	0.136 [0.127;0.146]	0.134 [0.124;0.144]
CH	0.092 [0.081;0.103]	0.083 [0.075;0.091]	0.063 [0.047;0.079]	0.080 [0.068;0.091]	0.090 [0.082;0.098]
CY	0.094 [0.085;0.103]	0.083 [0.073;0.094]	0.074 [0.058;0.090]	0.080 [0.066;0.094]	0.080 [0.065;0.095]
CZ	0.072 [0.065;0.079]	0.066 [0.059;0.073]	0.060 [0.051;0.069]	0.057 [0.048;0.066]	0.051 [0.044;0.058]
DE	0.070 [0.063;0.078]	0.059 [0.053;0.064]	0.047 [0.039;0.054]	0.070 [0.062;0.077]	0.079 [0.074;0.085]
DK	0.077 [0.046;0.108]	0.041 [0.030;0.052]	0.029 [0.018;0.040]	0.021 [0.011;0.031]	0.020 [0.015;0.026]
EE	0.111 [0.098;0.124]	0.102 [0.091;0.113]	0.074 [0.059;0.090]	0.097 [0.084;0.110]	0.101 [0.088;0.113]
EL	0.148 [0.130;0.165]	0.121 [0.110;0.132]	0.117 [0.099;0.134]	0.126 [0.111;0.142]	0.109 [0.094;0.124]
ES	0.142 [0.132;0.152]	0.120 [0.114;0.126]	0.089 [0.069;0.109]	0.128 [0.122;0.135]	0.120 [0.105;0.135]
FI	0.069 [0.049;0.088]	0.052 [0.041;0.062]	0.048 [0.032;0.063]	0.020 [0.009;0.031]	0.028 [0.021;0.034]
FR	0.086 [0.080;0.092]	0.086 [0.080;0.093]	0.072 [0.062;0.081]	0.090 [0.082;0.099]	0.098 [0.092;0.104]
HR	0.131 [0.117;0.146]	0.088 [0.080;0.097]	0.076 [0.064;0.088]	0.082 [0.070;0.095]	0.076 [0.066;0.087]
HU	0.110 [0.104;0.116]	0.103 [0.098;0.109]	0.095 [0.087;0.104]	0.113 [0.108;0.119]	0.108 [0.102;0.114]
IE	0.105 [0.092;0.118]	0.097 [0.087;0.108]	0.048 [0.029;0.068]	0.084 [0.070;0.099]	0.078 [0.069;0.087]
IS	0.067 [0.029;0.104]	0.032 [0.021;0.043]	0.030 [0.017;0.042]	0.012 [0.004;0.021]	0.016 [0.010;0.022]
IT	0.121 [0.113;0.130]	0.091 [0.086;0.095]	0.080 [0.068;0.091]	0.108 [0.102;0.113]	0.097 [0.090;0.104]
LT	0.095 [0.079;0.110]	0.067 [0.058;0.077]	0.059 [0.048;0.070]	0.069 [0.053;0.085]	0.067 [0.055;0.080]
LU	0.134 [0.125;0.143]	0.121 [0.114;0.127]	0.090 [0.072;0.109]	0.133 [0.125;0.140]	0.136 [0.130;0.142]
LV	0.134 [0.119;0.148]	0.110 [0.100;0.120]	0.095 [0.079;0.112]	0.110 [0.097;0.124]	0.111 [0.100;0.122]
MT	0.087 [0.075;0.099]	0.080 [0.071;0.089]	0.057 [0.047;0.067]	0.071 [0.059;0.083]	0.072 [0.062;0.082]
NL	0.066 [0.050;0.082]	0.053 [0.047;0.059]	0.041 [0.029;0.053]	0.028 [0.020;0.037]	0.019 [0.015;0.024]
NO	0.048 [0.032;0.064]	0.041 [0.031;0.050]	0.030 [0.019;0.041]	0.020 [0.012;0.028]	0.023 [0.018;0.029]
PL	0.111 [0.104;0.118]	0.097 [0.091;0.104]	0.095 [0.088;0.102]	0.102 [0.095;0.109]	0.099 [0.092;0.106]
PT	0.138 [0.128;0.148]	0.124 [0.113;0.134]	0.116 [0.102;0.129]	0.136 [0.124;0.149]	0.127 [0.114;0.140]
RO	0.170 [0.158;0.182]	0.104 [0.094;0.114]	0.119 [0.105;0.134]	0.120 [0.109;0.132]	0.111 [0.100;0.122]
SE	0.118 [0.037;0.199]	0.060 [0.043;0.078]	0.025 [0.007;0.043]	0.025 [0.016;0.033]	0.031 [0.025;0.038]
SI	0.077 [0.069;0.085]	0.073 [0.066;0.080]	0.059 [0.051;0.067]	0.032 [0.024;0.039]	0.036 [0.032;0.040]
SK	0.063 [0.055;0.071]	0.051 [0.045;0.057]	0.042 [0.033;0.051]	0.050 [0.041;0.058]	0.046 [0.039;0.053]
UK	0.101 [0.087;0.115]	0.090 [0.080;0.099]	0.062 [0.042;0.082]	0.071 [0.056;0.087]	0.079 [0.071;0.087]

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

Note: Inequality of opportunity is measured by the Gini coefficient in the estimated counterfactual distribution $\hat{y}^C$. $\hat{y}^C$ is constructed by the respective estimation approach indicated in the table header. 95% confidence intervals are derived based on 200 bootstrapped re-samples using the normal approximation method.

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

Paolo Brunori[*], Paul Hufe[†], Daniel Gerszon Mahler[‡]

Supplementary Material

[*]University of Florence, paolo.brunori@unifi.it.

[†]Corresponding author: University of Munich and ifo Institute, ifo Center for Macroeconomics and Surveys, Poschingerstr. 5, 81679 Munich, paul.hufe@econ.lmu.de.

[‡]The World Bank, dmahler@worldbank.org.

A DESCRIPTIVE STATISTICS

Table A.1: Descriptive Statistics (Individual and Household)

		Birth Area		Parents in HH		HH Composition			
Country	Male	Native	EU	Both	None	Adults	Working Ad.	Children	Home Owner
AT	0.501	0.790	0.070	0.856	0.017	2.730	1.760	2.600	0.585
BE	0.498	0.824	0.076	0.855	0.019	2.380	1.590	2.780	0.750
BG	0.500	0.994	0.001	0.904	0.012	2.440	2.010	2.070	0.910
CH	0.505	0.684	0.197	0.837	0.017	2.550	1.900	2.530	0.546
CY	0.525	0.787	0.096	0.900	0.015	2.640	1.670	2.700	0.784
CZ	0.508	0.964	0.026	0.851	0.013	2.090	1.920	2.240	0.597
DE	0.496	0.868	0.000	0.830	0.020	2.240	1.680	2.320	0.499
DK	0.505	0.923	0.026	0.809	0.027	2.220	2.310	2.240	0.736
EE	0.525	0.847	0.000	0.756	0.011	2.100	1.800	2.090	0.859
EL	0.498	0.890	0.025	0.931	0.019	2.310	1.560	2.330	0.834
ES	0.495	0.834	0.051	0.893	0.012	2.880	2.110	2.430	0.819
FI	0.499	0.954	0.018	0.829	0.016	2.360	1.750	2.300	0.772
FR	0.509	0.885	0.036	0.820	0.022	2.470	1.660	1.750	0.630
HR	0.501	0.875	0.017	0.874	0.020	2.560	1.350	2.310	0.902
HU	0.517	0.988	0.008	0.844	0.041	2.140	1.750	2.270	0.830
IE	0.524	0.783	0.149	0.893	0.078	3.170	3.200	3.200	0.727
IS	0.507	0.920	0.042	0.899	0.012	2.420	1.900	2.630	0.893
IT	0.502	0.880	0.040	0.901	0.011	2.590	1.620	2.410	0.685
LT	0.521	0.939	0.004	0.846	0.016	2.320	2.020	2.460	0.698
LU	0.499	0.480	0.401	0.868	0.020	2.530	1.640	2.710	0.734
LV	0.520	0.865	0.000	0.763	0.012	1.970	1.760	2.280	0.455
MT	0.497	0.944	0.000	0.932	0.020	3.020	1.840	2.680	0.576
NL	0.509	0.903	0.020	0.882	0.016	2.100	1.540	3.250	0.575
NO	0.511	0.907	0.041	0.913	0.014	2.020	1.760	1.870	0.922
PL	0.504	0.999	0.000	0.889	0.015	2.700	1.960	2.440	0.644
PT	0.506	0.906	0.022	0.854	0.017	2.680	2.230	2.680	0.544
RO	0.506	0.999	0.000	0.919	0.009	2.770	1.900	2.270	0.861
SE	0.493	0.846	0.050	0.820	0.035	2.070	1.780	2.350	0.757
SI	0.496	0.876	0.000	0.855	0.019	2.530	1.770	2.200	0.746
SK	0.519	0.987	0.010	0.920	0.010	2.520	2.080	2.340	0.694
UK	0.507	0.848	0.042	0.825	0.024	2.340	2.240	2.410	0.649

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

Table A.2: 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.
AT	0.743	0.093	0.777	0.068	0.398	0.421	0.135	0.714	0.215	0.003	0.001	0.043	0.046	0.064	0.051	0.138	0.145	0.284	0.063	0.085	0.338
BE	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
BG	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
CH	0.588	0.286	0.603	0.280	0.227	0.487	0.151	0.653	0.292	0.001	0.000	0.086	0.131	0.140	0.057	0.060	0.111	0.223	0.077	0.054	0.397
CY	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
CZ	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
DE	0.800	0.200	0.855	0.145	0.125	0.496	0.213	0.819	0.123	0.008	0.001	0.046	0.104	0.158	0.051	0.061	0.059	0.266	0.154	0.040	0.299
DK	0.935	0.025	0.970	0.020	0.368	0.418	0.214	0.708	0.272	0.004	0.001	0.111	0.122	0.070	0.043	0.103	0.160	0.288	0.072	0.009	0.447
EE	0.603	0.270	0.637	0.233	0.300	0.338	0.165	0.823	0.006	0.003	0.002	0.076	0.092	0.053	0.014	0.013	0.034	0.221	0.253	0.053	0.153
EL	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
ES	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
FI	0.827	0.007	0.827	0.007	0.491	0.182	0.162	0.592	0.209	0.016	0.001	0.041	0.089	0.085	0.016	0.046	0.135	0.146	0.138	0.044	
FR	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
HR	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
HU	0.962	0.017	0.969	0.012	0.599	0.241	0.087	0.892	0.043	0.001	0.001	0.037	0.060	0.052	0.017	0.053	0.094	0.279	0.193	0.137	0.117
IE	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
IS	0.918	0.050	0.923	0.044	0.334	0.486	0.139	0.638	0.332	0.001	0.000	0.115	0.121	0.076	0.024	0.094	0.180	0.220	0.094	0.042	0.570
IT	0.823	0.022	0.827	0.020	0.708	0.136	0.038	0.614	0.244	0.016	0.004	0.054	0.040	0.074	0.057	0.068	0.099	0.227	0.105	0.118	0.199
LT	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
LU	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
LV	0.572	0.248	0.642	0.165	0.381	0.297	0.098	0.767	0.005	0.002	0.003	0.036	0.083	0.037	0.010	0.019	0.069	0.199	0.218	0.083	0.070
MT	0.952	0.041	0.953	0.040	0.561	0.180	0.059	0.717	0.214	0.013	0.001	0.062	0.046	0.106	0.045	0.141	0.050	0.244	0.099	0.106	0.225
NL	0.829	0.028	0.888	0.022	0.376	0.285	0.198	0.726	0.173	0.006	0.006	0.087	0.124	0.155	0.051	0.069	0.086	0.200	0.079	0.031	0.310
NO	0.897	0.046	0.908	0.041	0.328	0.390	0.278	0.712	0.255	0.002	0.001	0.116	0.110	0.167	0.029	0.057	0.111	0.227	0.100	0.032	0.285
PL	0.955	0.012	0.980	0.003	0.462	0.448	0.070	0.701	0.238	0.002	0.001	0.036	0.044	0.053	0.025	0.042	0.237	0.254	0.157	0.078	0.111
PT	0.932	0.006	0.945	0.006	0.700	0.031	0.031	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
RO	0.938	0.001	0.939	0.001	0.726	0.088	0.030	0.642	0.237	0.004	0.013	0.004	0.040	0.034	0.016	0.018	0.253	0.249	0.121	0.104	0.045
SE	0.945	0.022	0.851	0.061	0.422	0.350	0.182	0.745	0.211	0.002	0.001	0.043	0.118	0.067	0.031	0.092	0.086	0.230	0.108	0.019	0.337
SI	0.769	0.200	0.000	0.000	0.684	0.166	0.085	0.773	0.099	0.013	0.011	0.024	0.052	0.100	0.037	0.052	0.089	0.257	0.080	0.173	0.242
SK	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
UK	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 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". Compare also to Table 1.

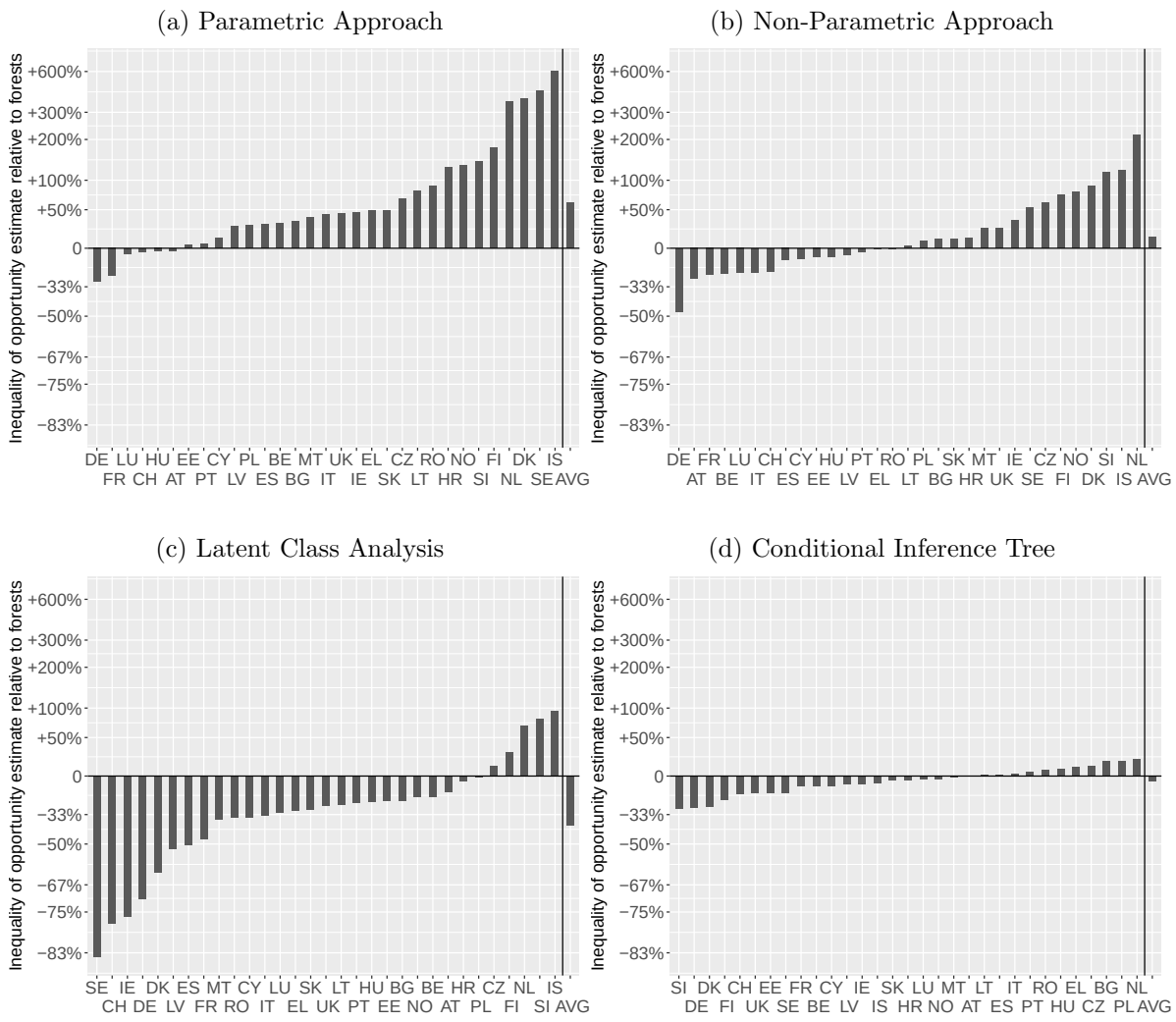
Table A.3: Descriptive Statistics (Mothers)

	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.
AT	0.740	0.096	0.789	0.065	0.587	0.328	0.041	0.369	0.169	0.002	0.435	0.009	0.024	0.009	0.071	0.154	0.128	0.045	0.010	0.087	0.092
BE	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
BG	0.931	0.003	0.981	0.002	0.464	0.357	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	0.152	0.030
CH	0.567	0.307	0.599	0.286	0.410	0.399	0.057	0.382	0.152	0.001	0.429	0.027	0.056	0.069	0.069	0.125	0.055	0.039	0.025	0.068	0.064
CY	0.804	0.080	0.812	0.091	0.684	0.162	0.057	0.325	0.166	0.001	0.492	0.001	0.045	0.020	0.037	0.067	0.036	0.022	0.042	0.220	0.048
CZ	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
DE	0.811	0.189	0.862	0.138	0.284	0.475	0.081	0.482	0.050	0.009	0.438	0.012	0.051	0.079	0.089	0.116	0.025	0.015	0.087	0.033	0.059
DK	0.922	0.029	0.935	0.023	0.531	0.283	0.186	0.630	0.069	0.006	0.274	0.020	0.103	0.095	0.123	0.225	0.035	0.052	0.026	0.001	0.122
EE	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
EL	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
ES	0.836	0.046	0.849	0.046	0.802	0.048	0.040	0.186	0.069	0.001	0.718	0.010	0.025	0.010	0.021	0.059	0.028	0.021	0.009	0.071	0.029
FI	0.826	0.007	0.933	0.006	0.559	0.238	0.145	0.658	0.204	0.019	0.055	0.012	0.126	0.091	0.122	0.145	0.048	0.046	0.057	0.202	
FR	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
HR	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
HU	0.964	0.016	0.980	0.012	0.655	0.243	0.053	0.729	0.022	0.001	0.216	0.014	0.049	0.063	0.113	0.118	0.061	0.075	0.087	0.167	0.044
IE	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
IS	0.905	0.059	0.924	0.046	0.626	0.275	0.075	0.598	0.102	0.001	0.278	0.030	0.095	0.045	0.109	0.180	0.064	0.028	0.013	0.130	0.149
IT	0.820	0.024	0.862	0.024	0.779	0.112	0.023	0.224	0.080	0.005	0.637	0.011	0.038	0.022	0.029	0.051	0.035	0.031	0.022	0.062	0.041
LT	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
LU	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
LV	0.585	0.234	0.793	0.182	0.414	0.399	0.125	0.891	0.003	0.002	0.064	0.031	0.138	0.084	0.098	0.121	0.085	0.093	0.023	0.221	0.074
MT	0.950	0.043	0.957	0.038	0.652	0.145	0.026	0.073	0.015	0.001	0.894	0.003	0.019	0.007	0.009	0.018	0.002	0.004	0.009	0.010	0.011
NL	0.829	0.027	0.907	0.023	0.532	0.288	0.087	0.282	0.056	0.003	0.628	0.010	0.050	0.038	0.052	0.089	0.016	0.011	0.008	0.060	0.037
NO	0.877	0.048	0.891	0.043	0.368	0.437	0.181	0.623	0.106	0.008	0.239	0.031	0.041	0.142	0.114	0.209	0.053	0.017	0.026	0.091	0.065
PL	0.957	0.010	0.990	0.004	0.524	0.410	0.057	0.518	0.261	0.008	0.174	0.018	0.057	0.053	0.071	0.096	0.262	0.080	0.018	0.118	0.050
PT	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	0.075	0.158	0.059	0.032	0.145	0.048
RO	0.936	0.001	0.939	0.001	0.728	0.112	0.020	0.370	0.219	0.005	0.300	0.001	0.034	0.024	0.026	0.050	0.218	0.076	0.040	0.080	0.010
SE	0.942	0.024	0.855	0.058	0.409	0.369	0.201	0.731	0.058	0.002	0.189	0.006	0.087	0.033	0.057	0.152	0.016	0.009	0.021	0.035	0.095
SI	0.791	0.178	0.000	0.000	0.752	0.148	0.058	0.578	0.071	0.005	0.296	0.006	0.047	0.093	0.085	0.090	0.061	0.066	0.006	0.193	0.089
SK	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
UK	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". Compare also to Table 1.

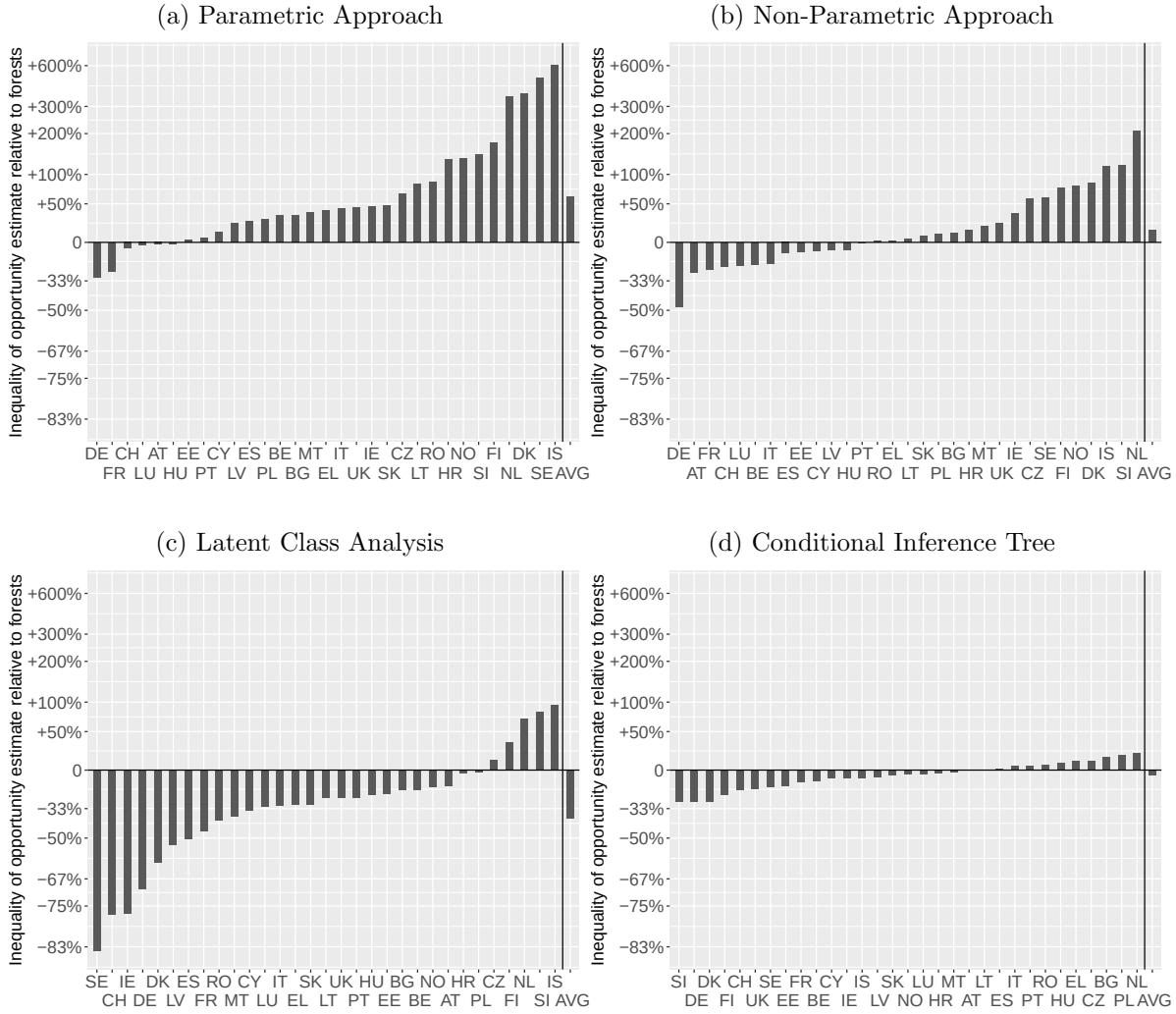
B ALTERNATIVE INEQUALITY INDIXES

Figure B.1: Correlation of Estimates by Method (GE(0))



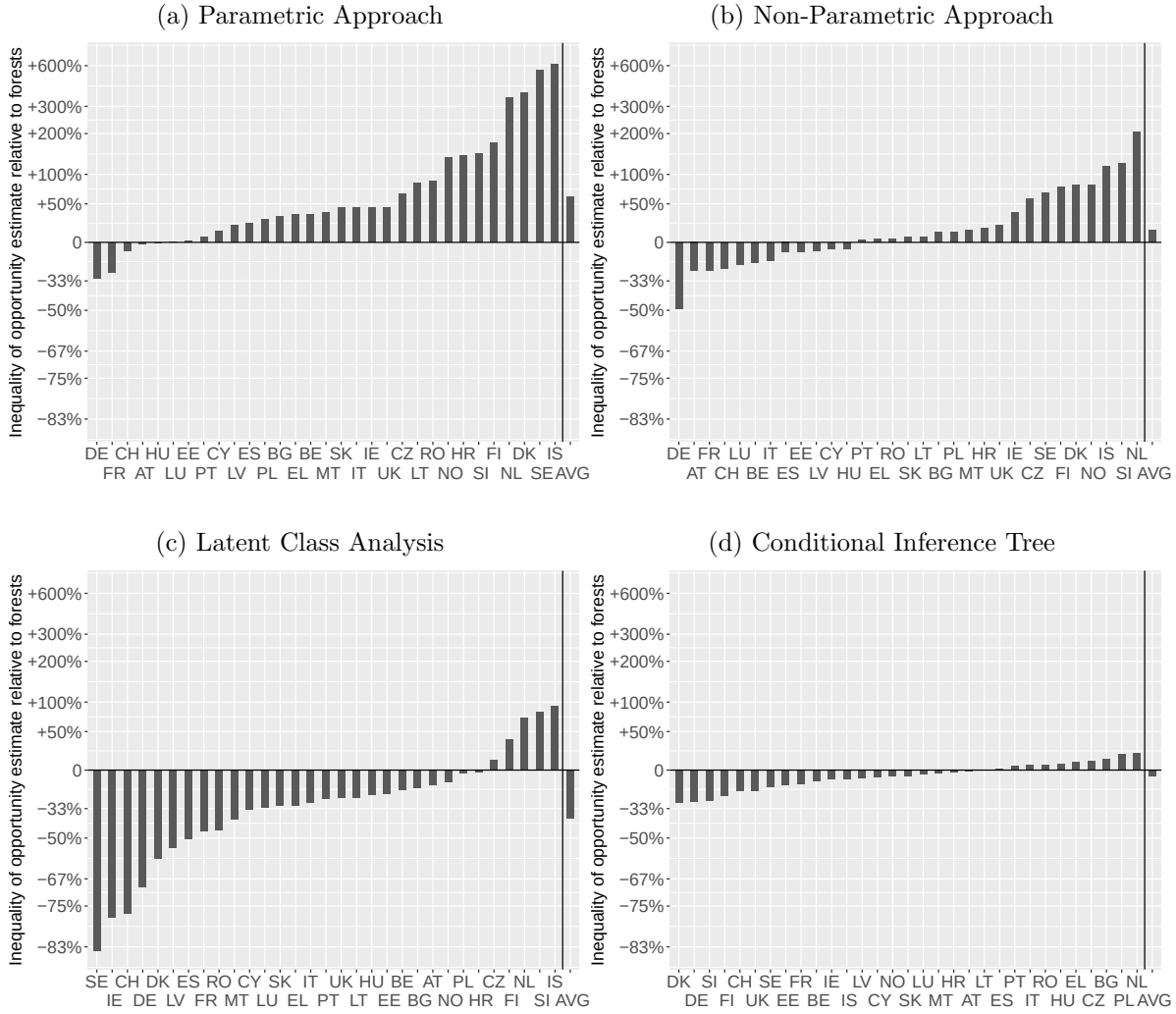
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: In each panel, the y-axis shows the inequality of opportunity estimate from the method in question divided by the inequality of opportunity estimate from forests, displayed on a logarithmic scale. Country-estimates above the black line indicate an overestimation of inequality of opportunity relative to the random forest benchmark. Reversely, country-estimates below the black line indicate an underestimation of inequality of opportunity relative to the random forest benchmark. For all methods inequality of opportunity is measured by the GE(0) index in the estimated counterfactual distribution $\hat{y}^C$. The figure is top (bottom) coded at +600% (−83%).

Figure B.2: Correlation of Estimates by Method (GE(1))



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: In each panel, the y-axis shows the inequality of opportunity estimate from the method in question divided by the inequality of opportunity estimate from forests, displayed on a logarithmic scale. Country-estimates above the black line indicate an overestimation of inequality of opportunity relative to the random forest benchmark. Reversely, country-estimates below the black line indicate an underestimation of inequality of opportunity relative to the random forest benchmark. For all methods inequality of opportunity is measured by the GE(1) index in the estimated counterfactual distribution $\hat{y}^C$. The figure is top (bottom) coded at +600% (−83%).

Figure B.3: Correlation of Estimates by Method (GE(2))

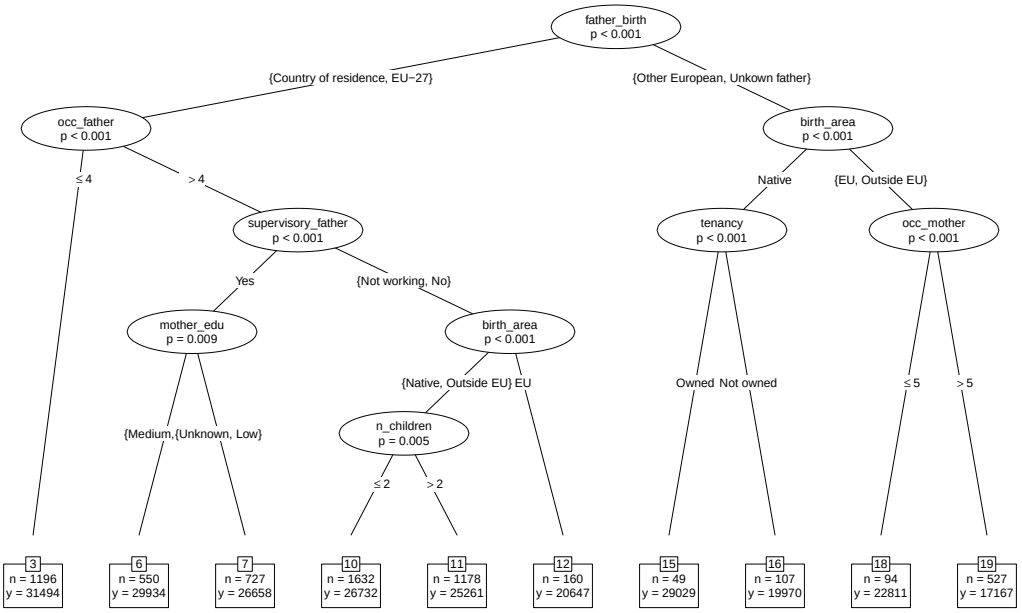


Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: In each panel, the y-axis shows the inequality of opportunity estimate from the method in question divided by the inequality of opportunity estimate from forests, displayed on a logarithmic scale. Country-estimates above the black line indicate an overestimation of inequality of opportunity relative to the random forest benchmark. Reversely, country-estimates below the black line indicate an underestimation of inequality of opportunity relative to the random forest benchmark. For all methods inequality of opportunity is measured by the GE(2) index in the estimated counterfactual distribution $\hat{y}^C$. The figure is top (bottom) coded at +600% (−83%).

C OPPORTUNITY STRUCTURES

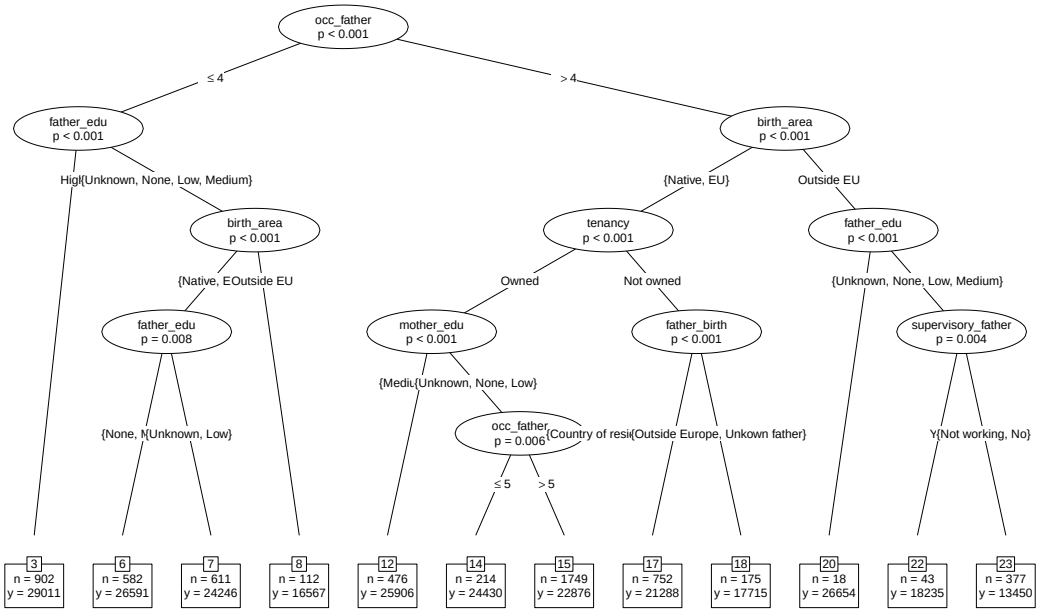
Trees

Figure C.4: Opportunity Tree (Austria)



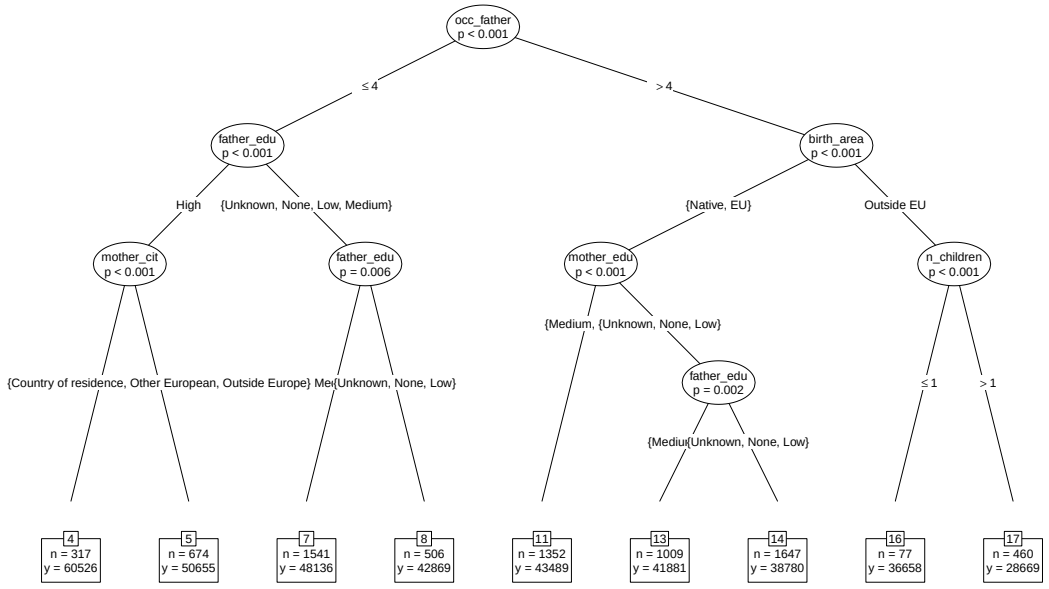
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.5: Opportunity Tree (Belgium)



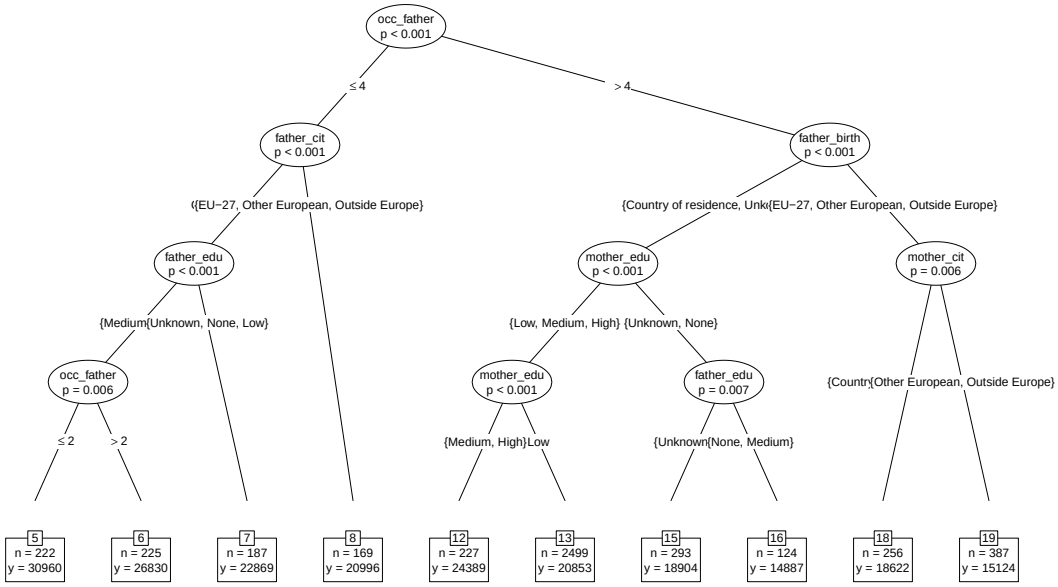
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.6: Opportunity Tree (Switzerland)



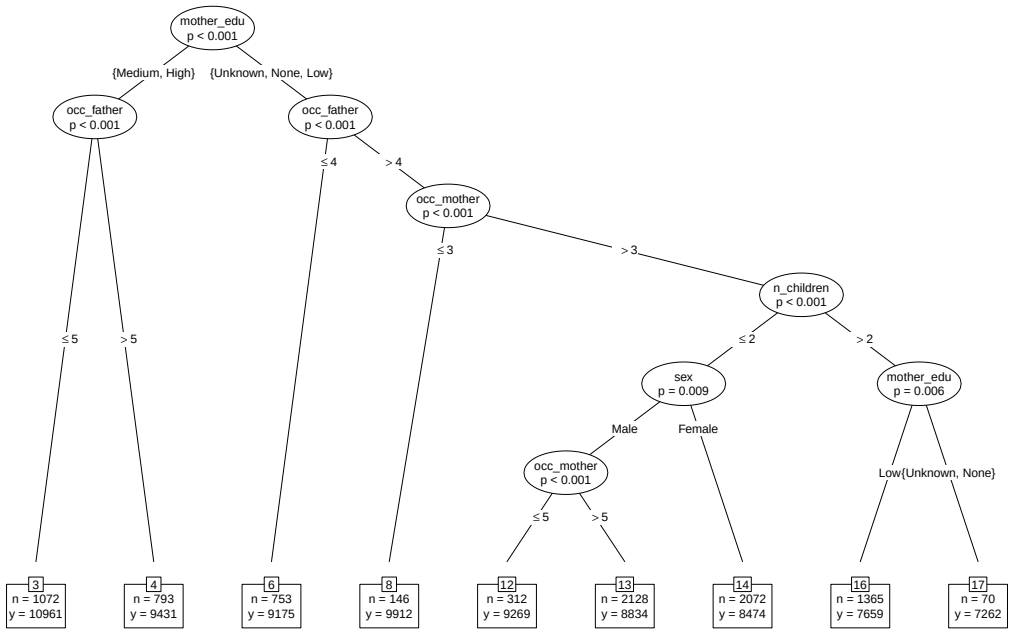
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.7: Opportunity Tree (Cyprus)



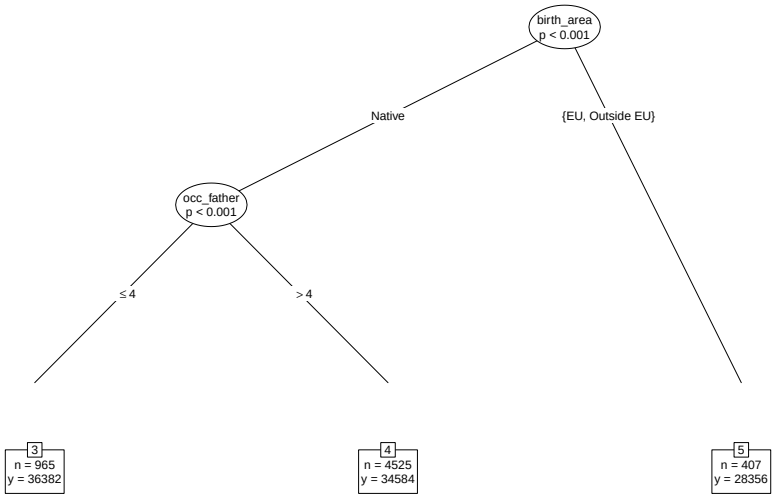
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.8: Opportunity Tree (Czech Republic)



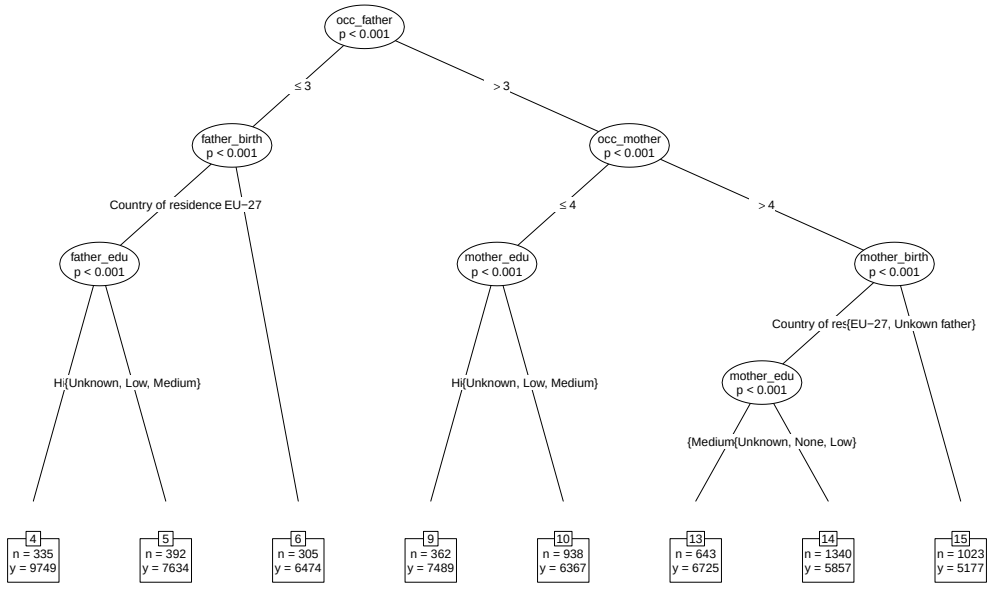
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.9: Opportunity Tree (Denmark)



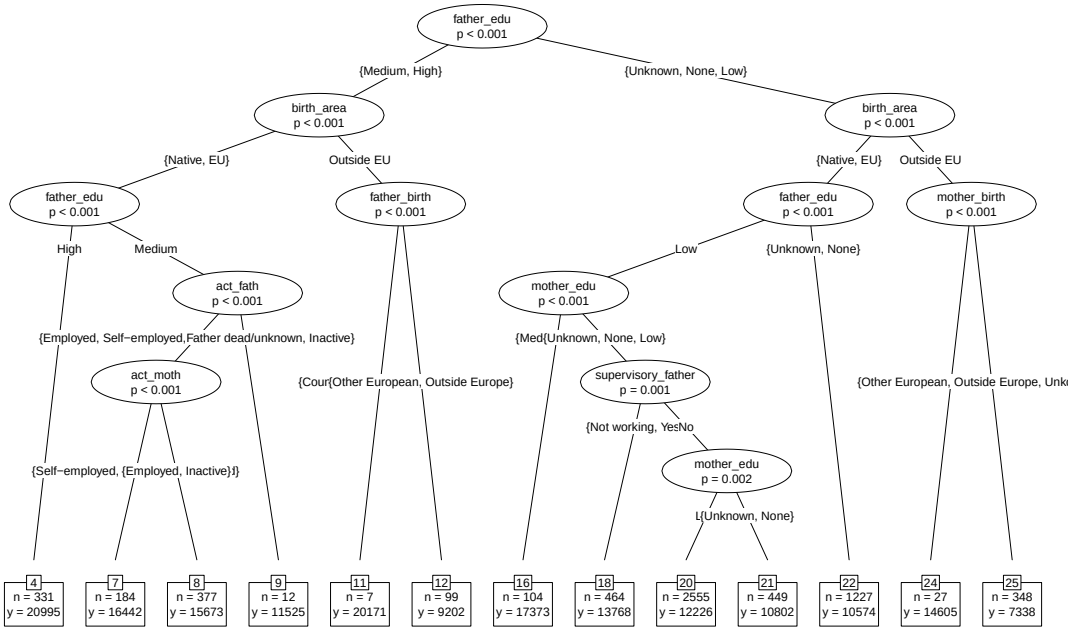
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.10: Opportunity Tree (Estonia)



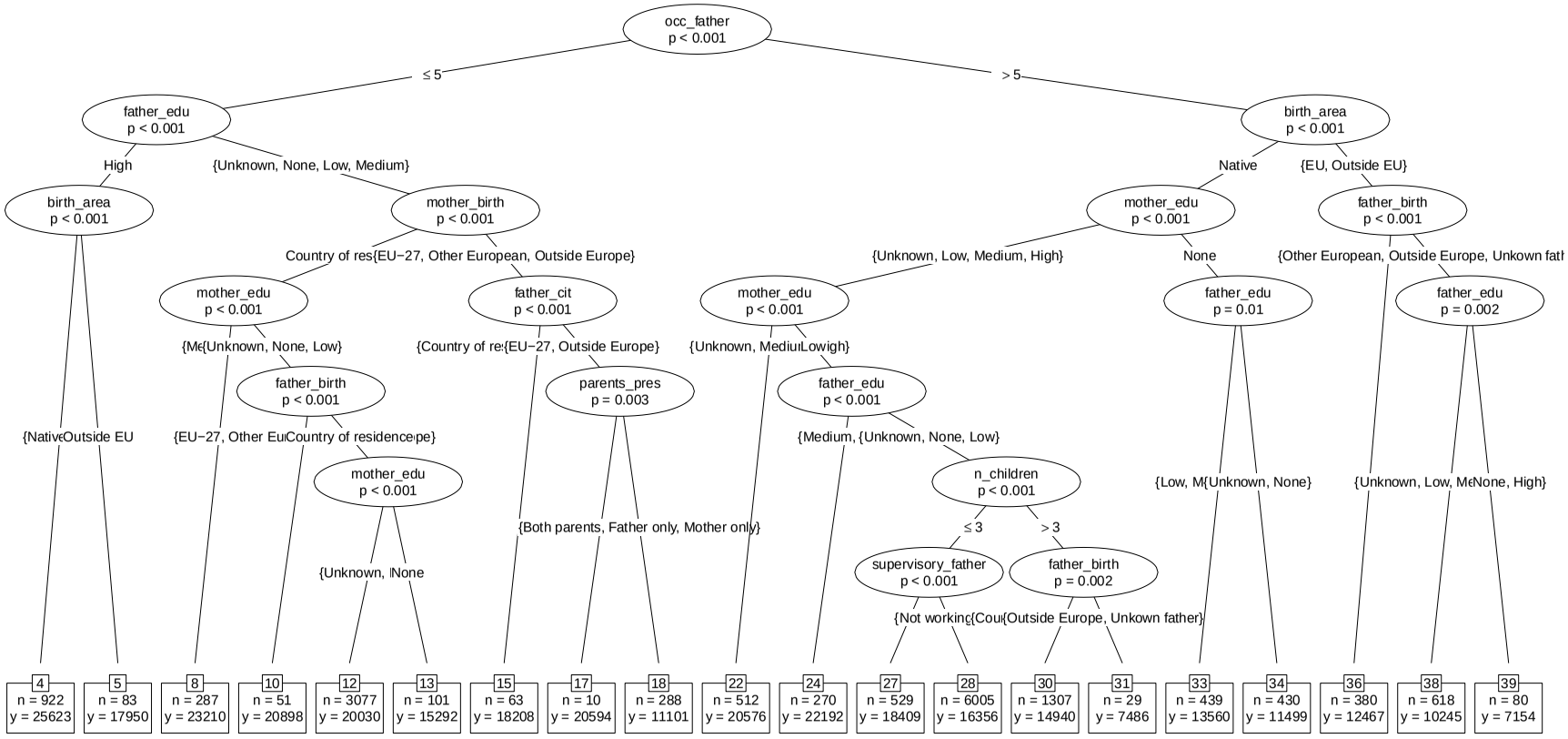
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.11: Opportunity Tree (Greece)



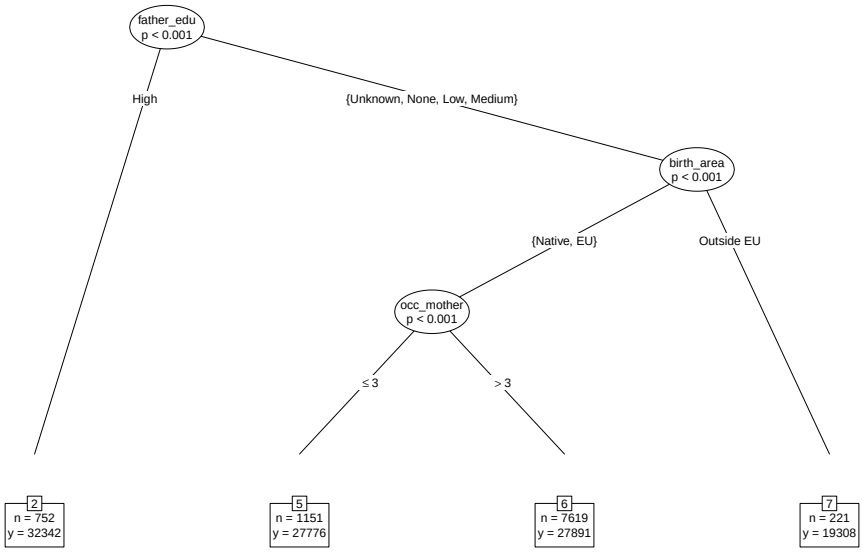
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.12: Opportunity Tree (Spain)



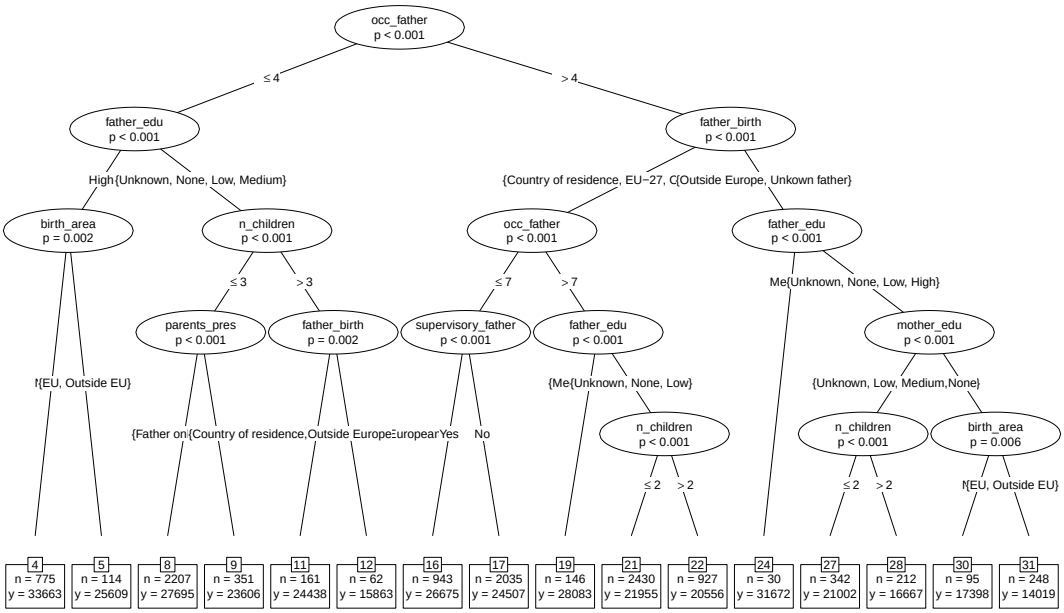
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.13: Opportunity Tree (Finland)



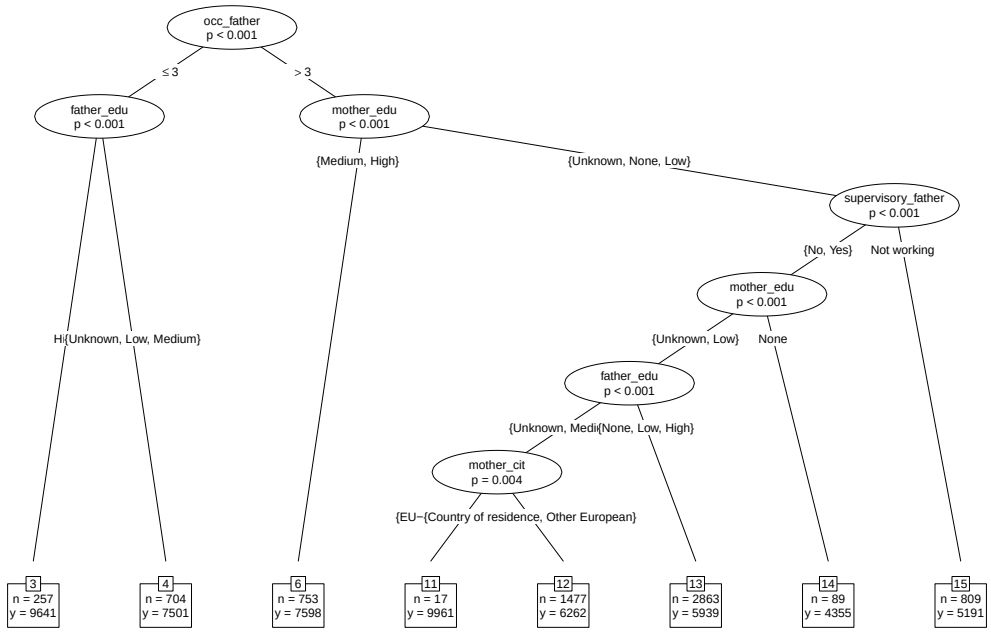
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.14: Opportunity Tree (France)



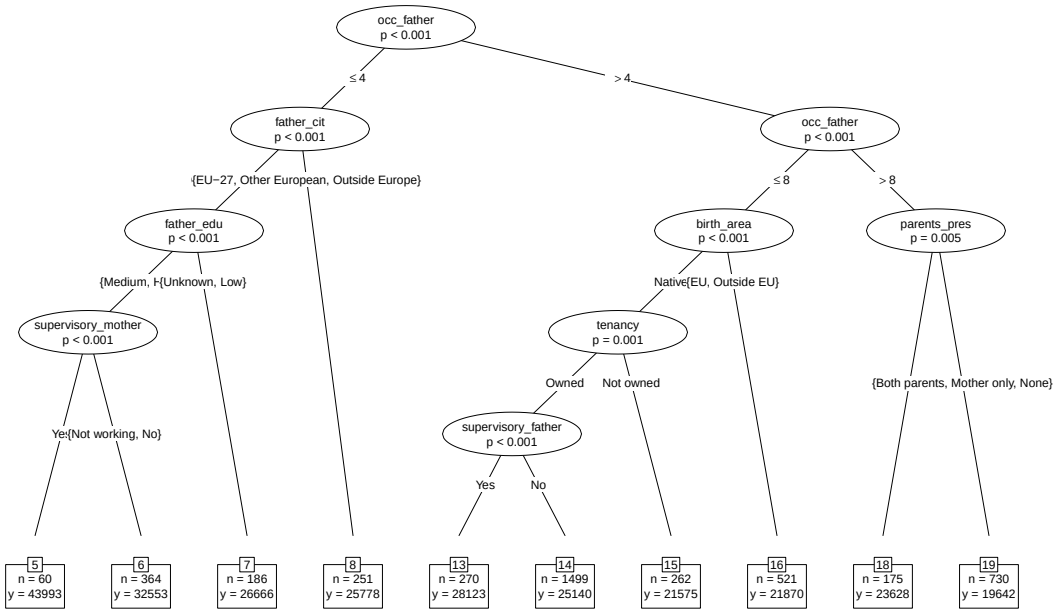
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.15: Opportunity Tree (Croatia)



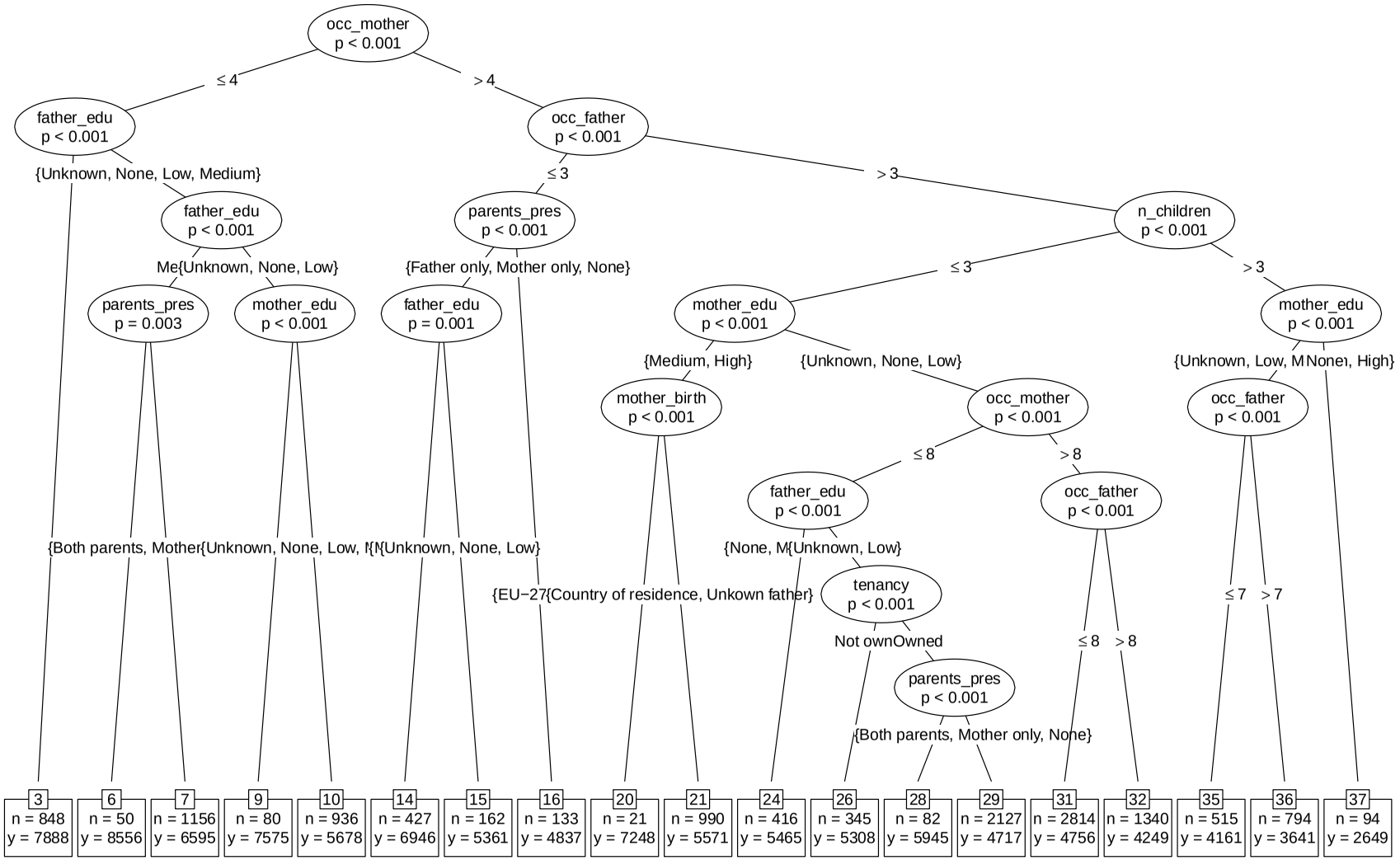
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.16: Opportunity Tree (Ireland)



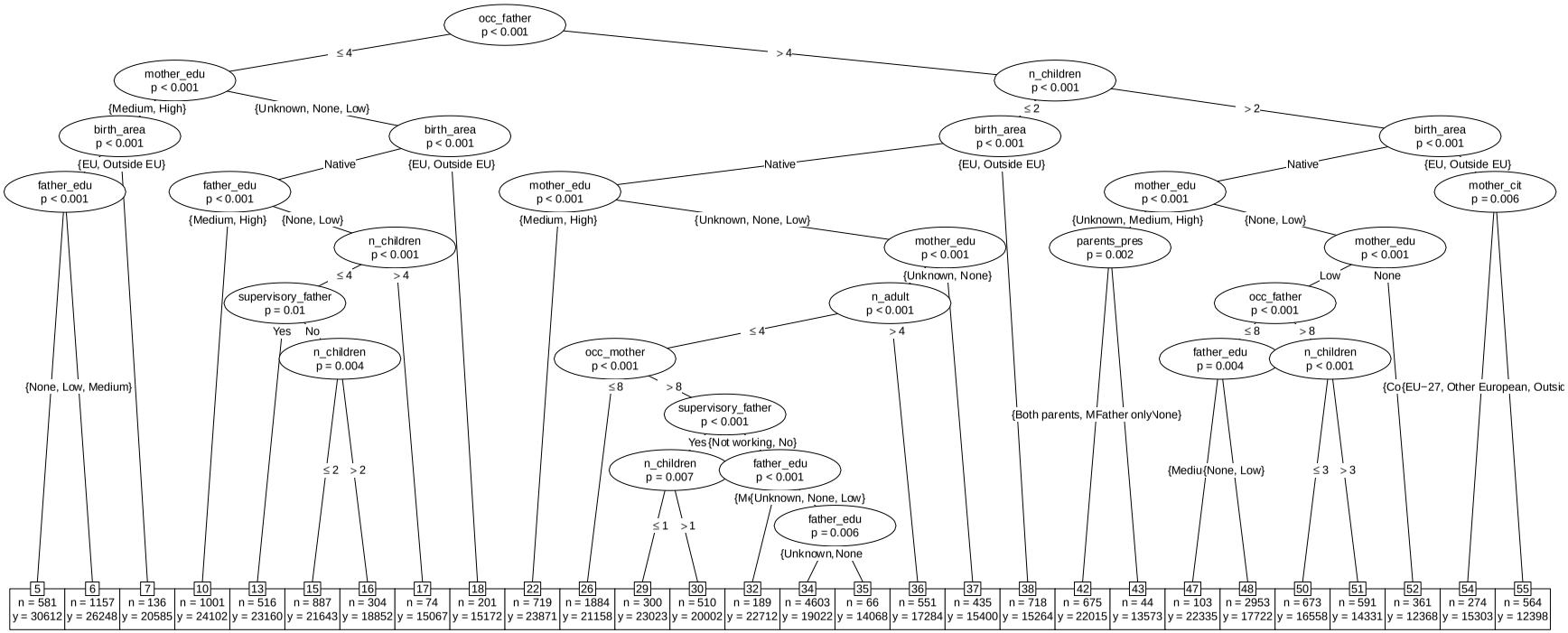
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.17: Opportunity Tree (Hungary)



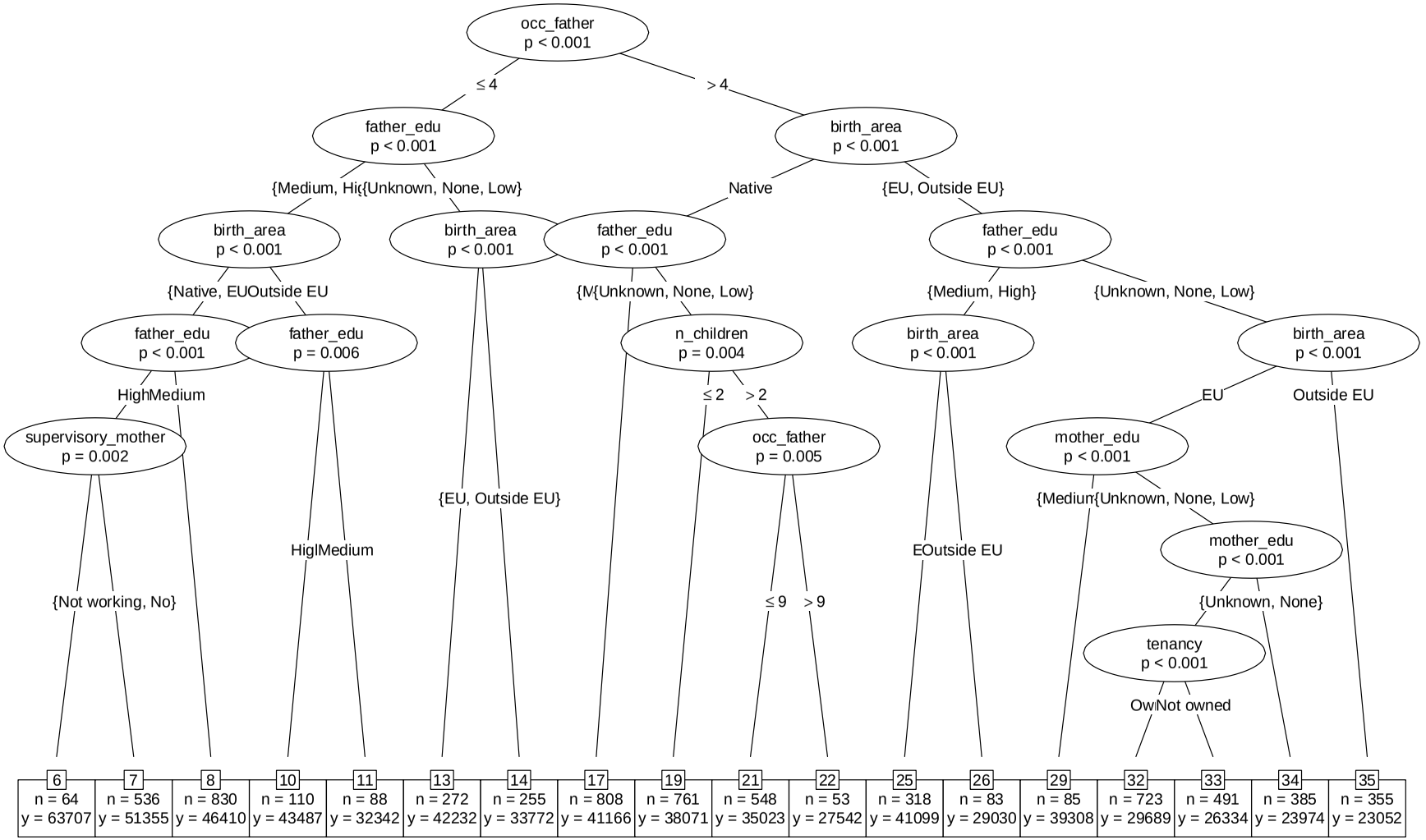
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.18: Opportunity Tree (Italy)



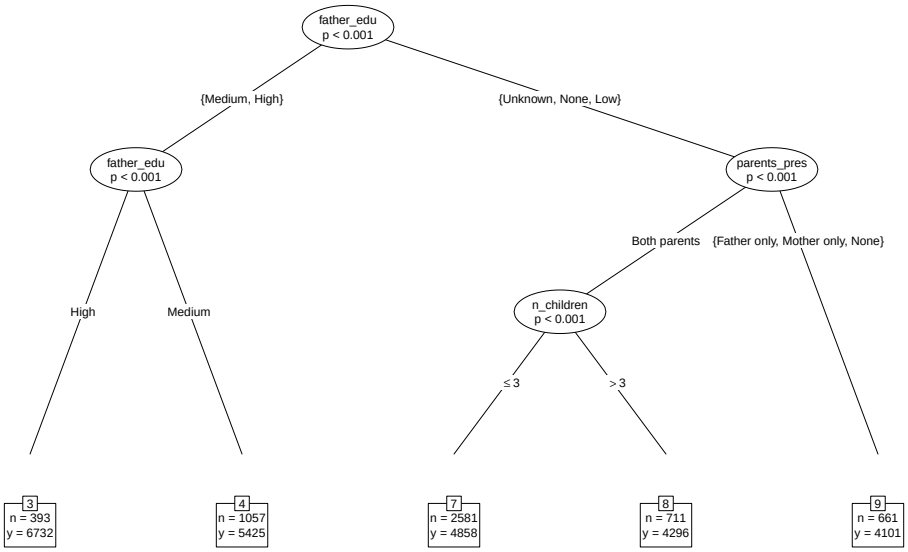
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.19: Opportunity Tree (Luxembourg)



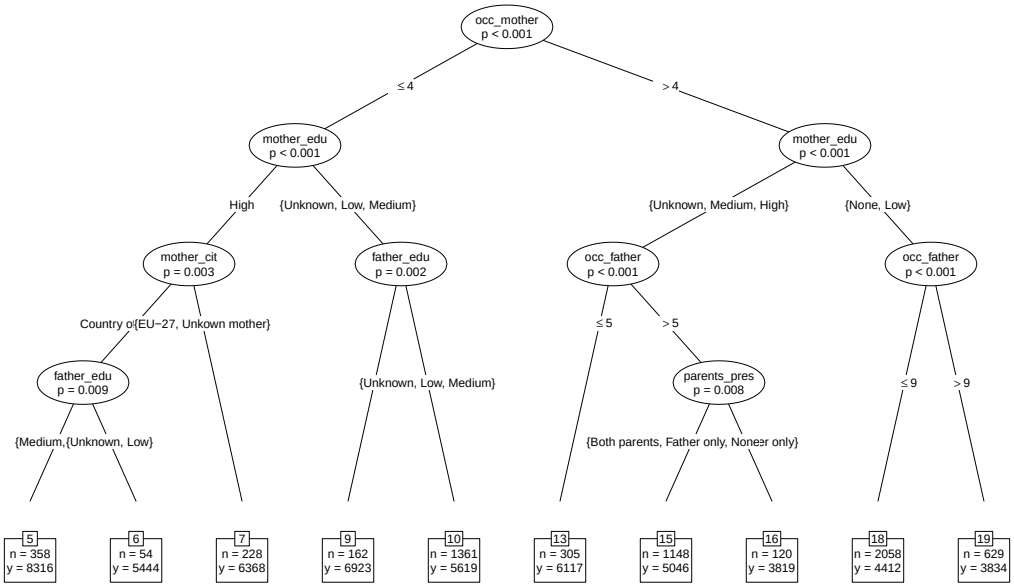
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.20: Opportunity Tree (Lithuania)



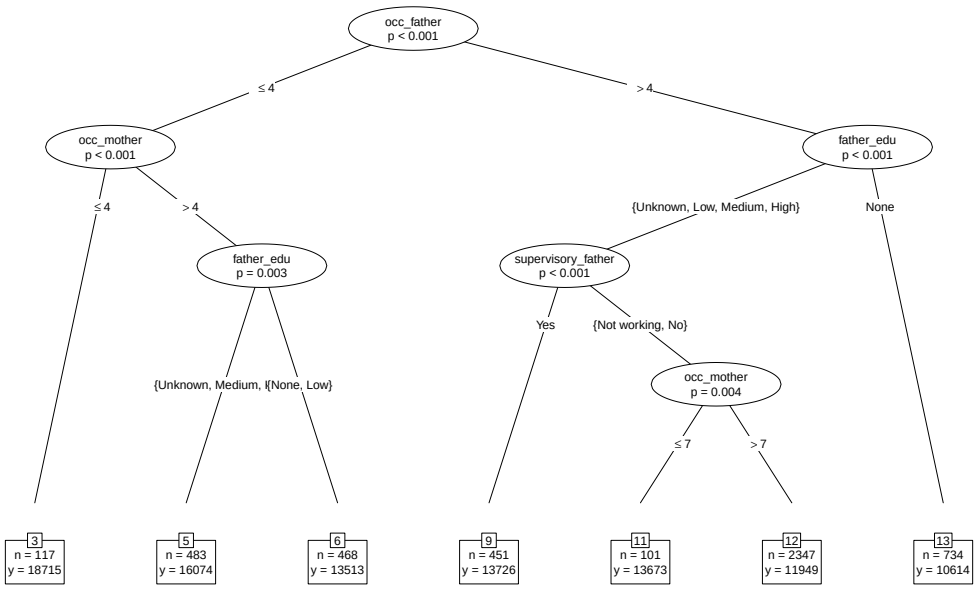
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.21: Opportunity Tree (Latvia)



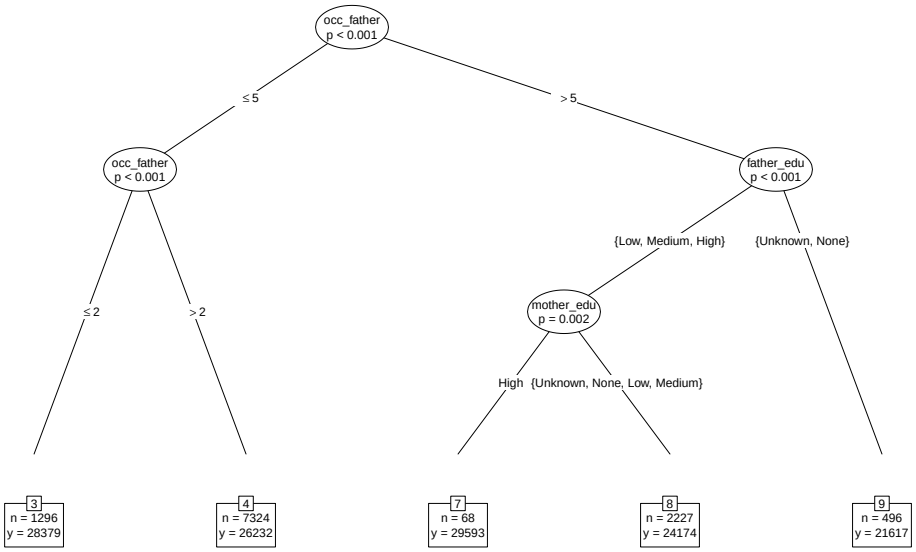
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.22: Opportunity Tree (Malta)



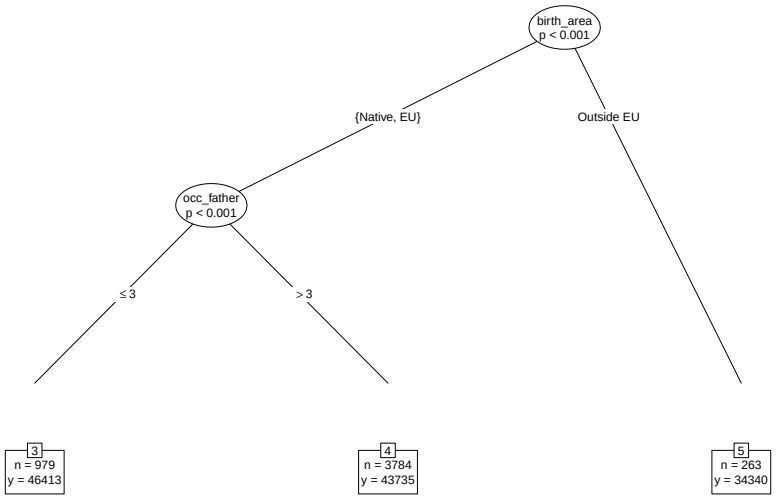
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.23: Opportunity Tree (Netherlands)



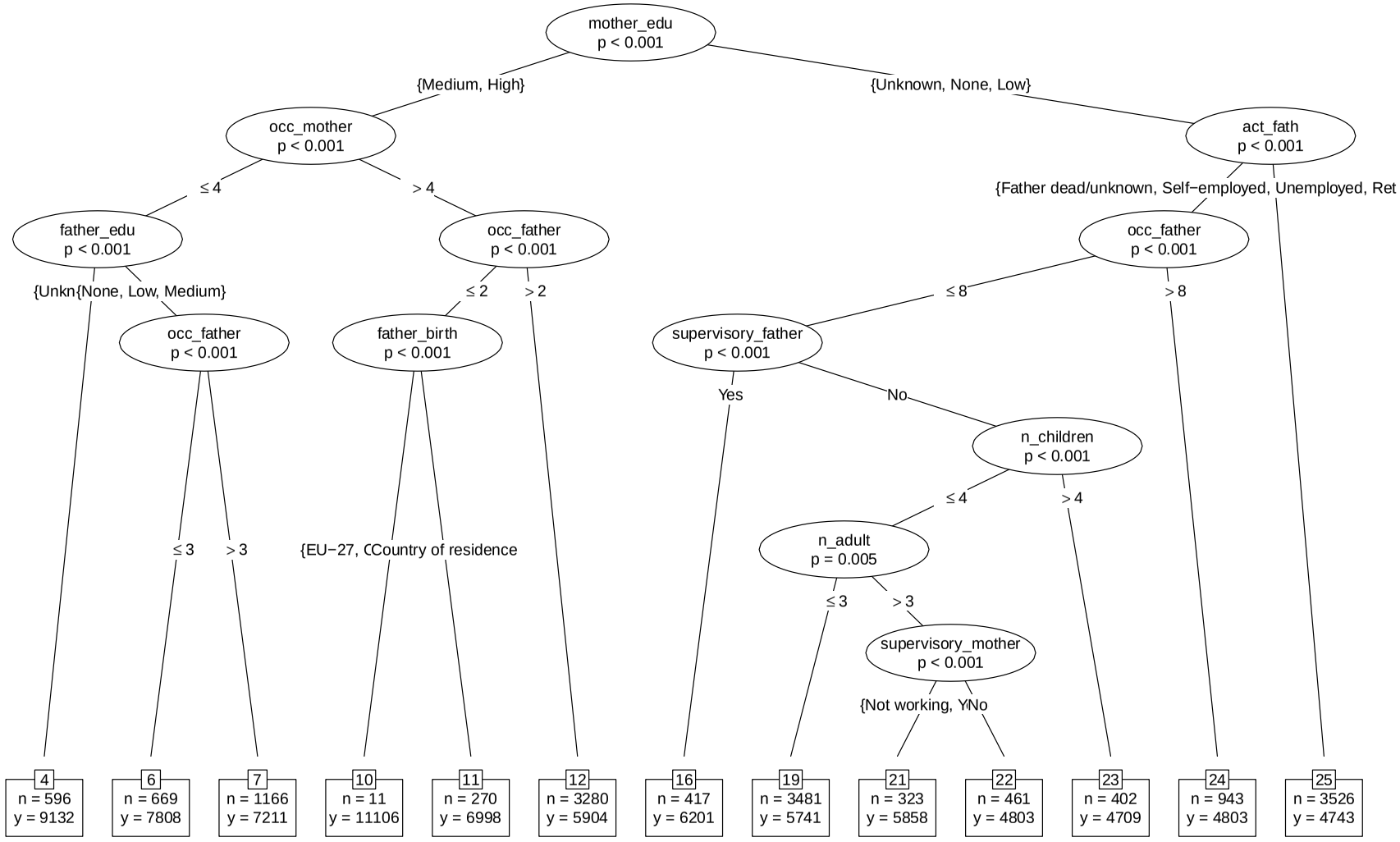
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.24: Opportunity Tree (Norway)



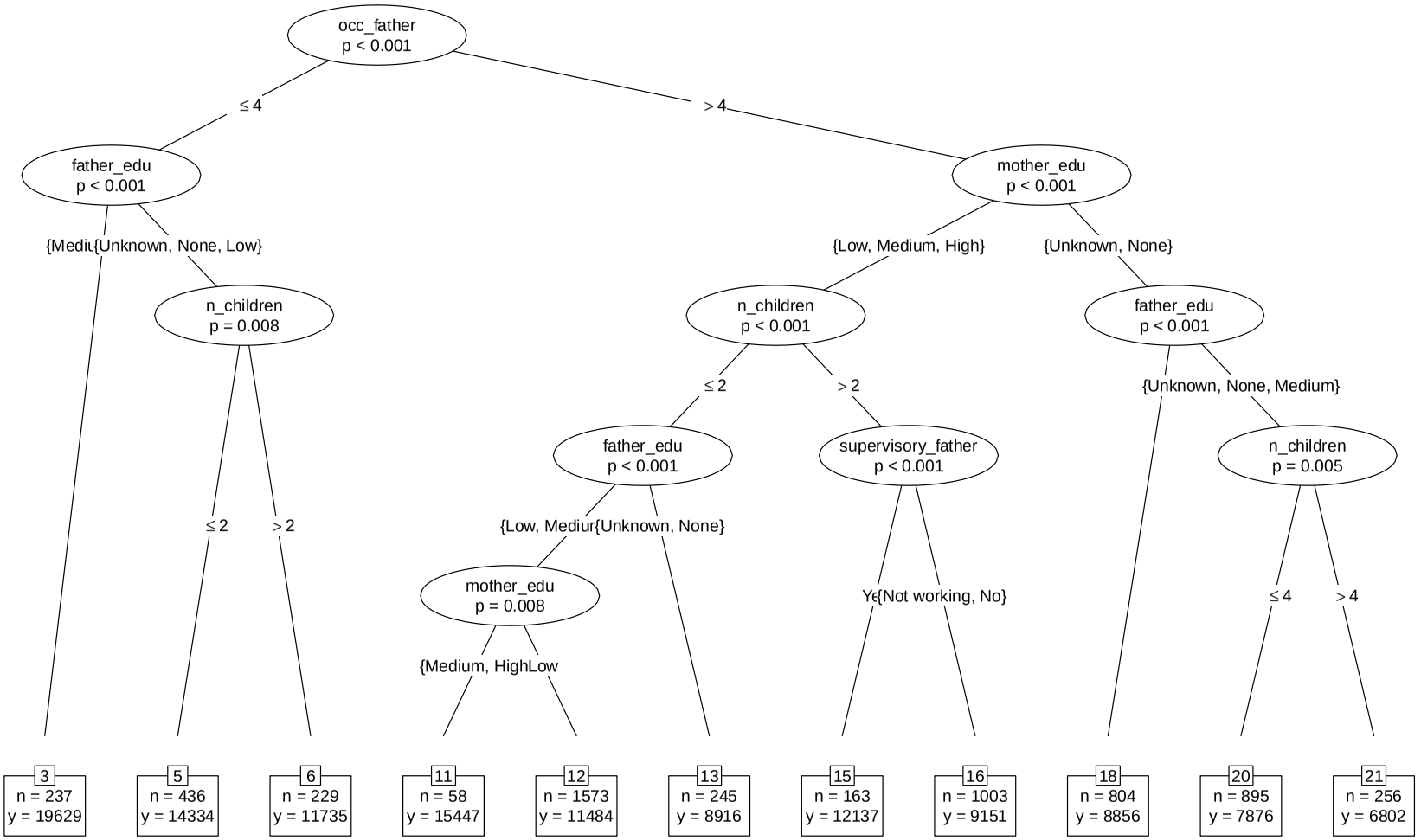
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.25: Opportunity Tree (Poland)



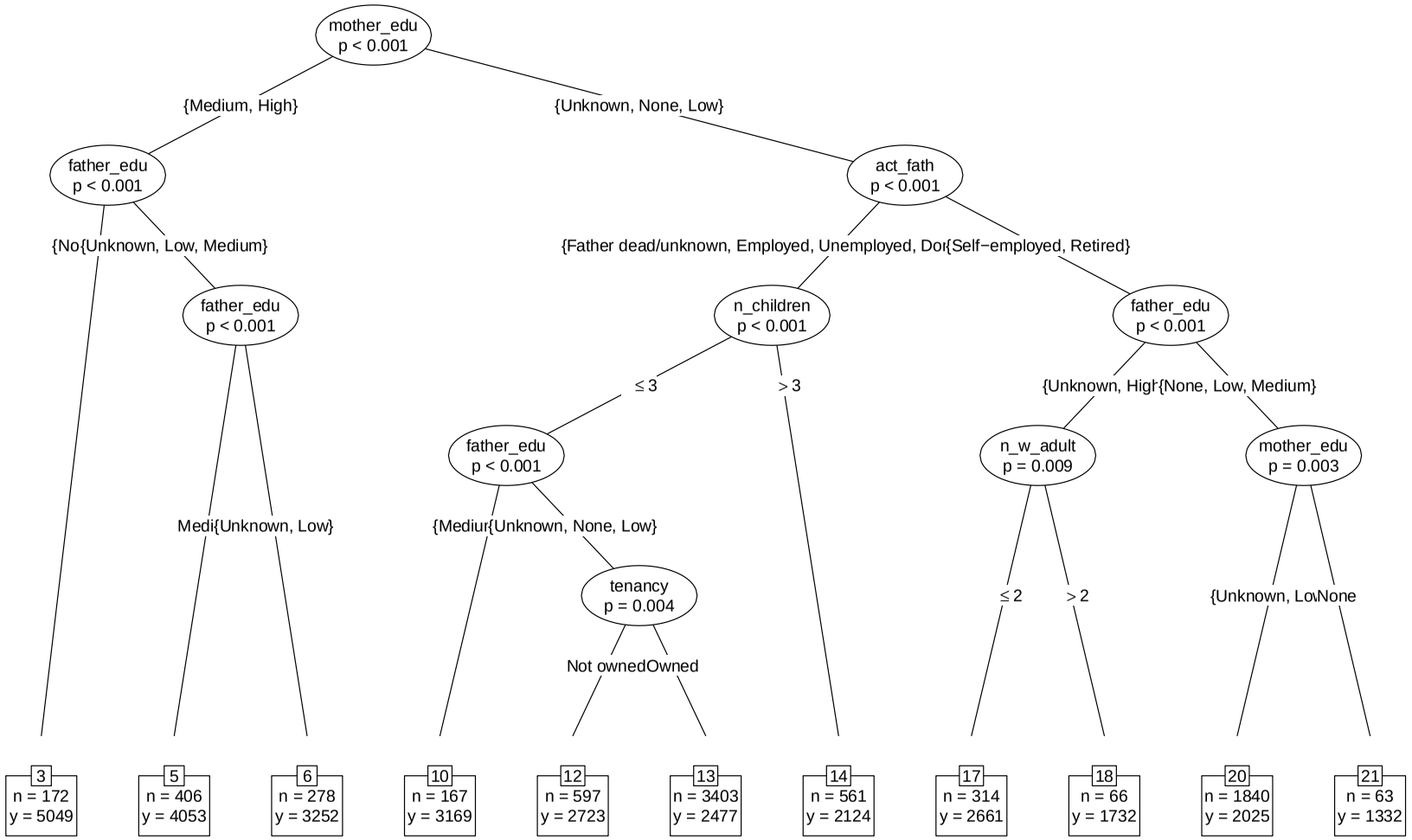
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.26: Opportunity Tree (Portgual)



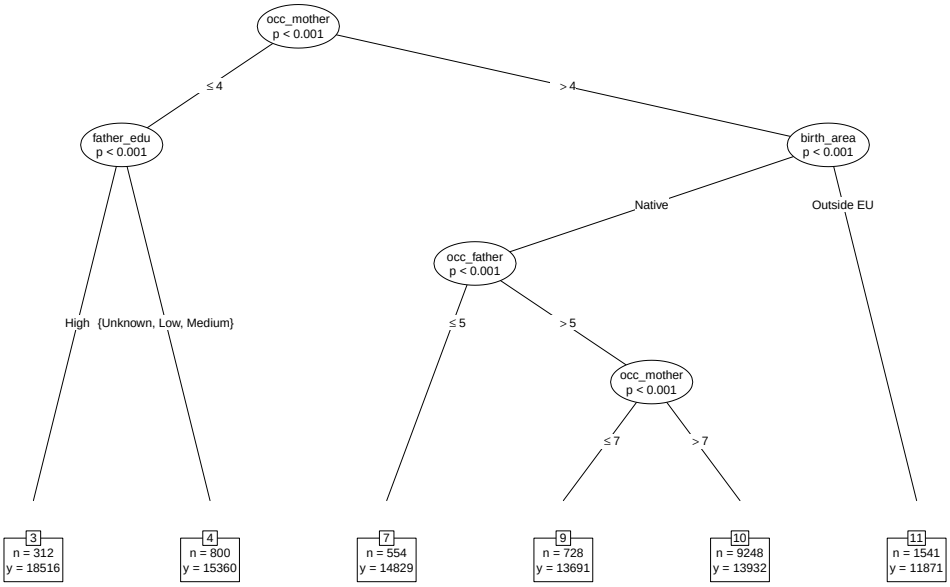
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.27: Opportunity Tree (Romania)



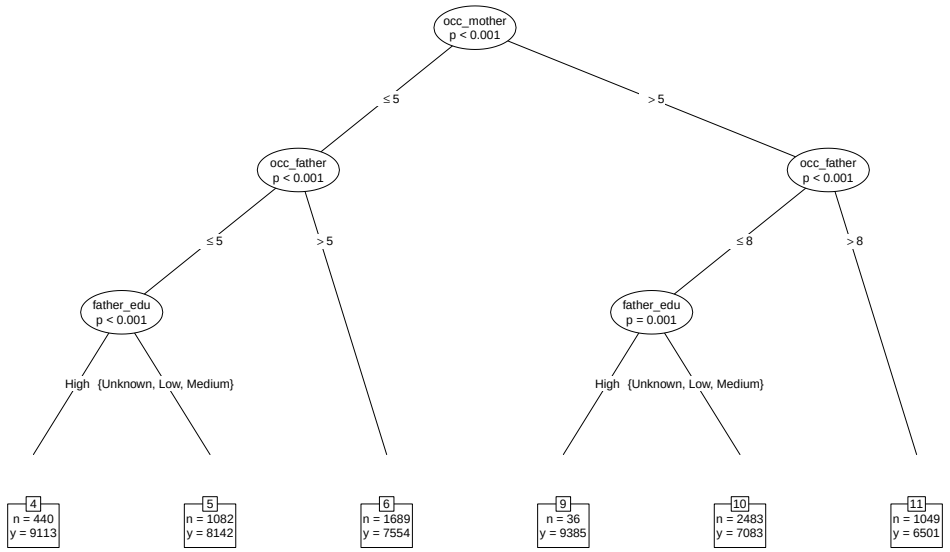
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.28: Opportunity Tree (Slovenia)



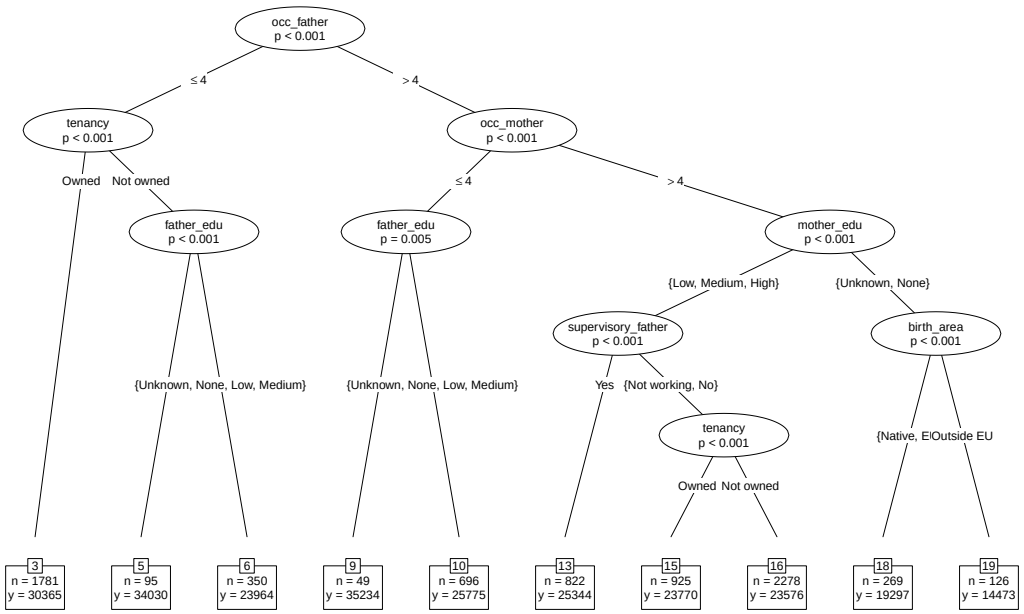
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.29: Opportunity Tree (Slovakia)



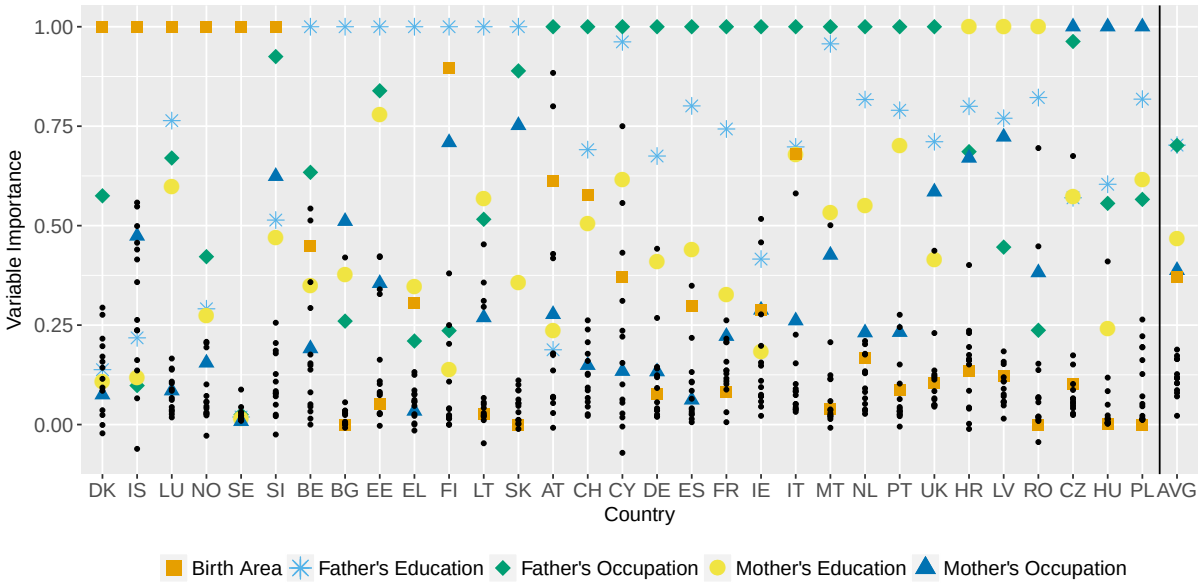
Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: The tree is constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1. Ellipses indicate splitting points, while 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 C.30: 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 II). The set of observed circumstances $\mathbf{\Omega}$ 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.

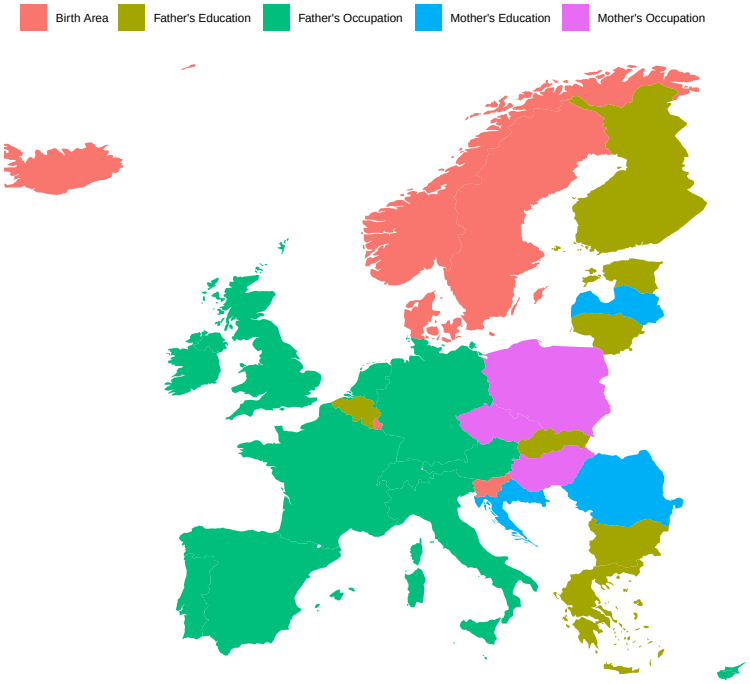
Figure C.31: Variable Importance Plot



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. The forests are constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1.

Figure C.32: Variable Importance Map



Data: EU-SILC 2011 cross-sectional (rev.5, June 2015).
Note: Each color shows the most important circumstance in each country. Variable importance is measured by the decrease in MSE^{OOB} after permuting ω^p such that it is orthogonal to y. The forests are constructed by the conditional inference algorithm (Section II). The set of observed circumstances $\mathbf{\Omega}$ used to construct the conditional inference tree is detailed in Table 1.