WPS8349

Policy Research Working Paper 8349

The Roots of Inequality

Estimating Inequality of Opportunity from Regression Trees

Paolo Brunori
Paul Hufe
Daniel Gerszon Mahler

WORLD BANK GROUP
Development Research Group
Poverty and Inequality Team
February 2018

Policy Research Working Paper 8349

Abstract

This paper proposes a set of new methods to estimate inequality of opportunity based on conditional inference regression trees. It illustrates how these methods represent a substantial improvement over existing empirical approaches to measure inequality of opportunity. First, the new methods minimize the risk of arbitrary and ad hoc model selection. Second, they provide a standardized way to trade off upward and downward biases in inequality of opportunity estimations. Finally, regression trees can be graphically represented; their structure is immediate to read and easy to understand. This will make the measurement of inequality of opportunity more easily comprehensible to a large audience. These advantages are illustrated by an empirical application based on the 2011 wave of the European Union Statistics on Income and Living Conditions.

This paper is a product of the Poverty and Inequality Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at dmahler@worldbank.org.

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

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

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

*We are grateful for comments received from participants during presentations held at the Institute for Social and Economic Research at the University of Essex, the Poverty and Applied Microeconomics Seminar at the World Bank, the Copenhagen Centre for Social Data Science at the University of Copenhagen, and the 13th Winter School on Social Choice Theory and Welfare at Canazei. Any errors remain our own.

†Corresponding Author: University of Florence, Dipartimento di Scienze per l'Economia e l'Impresa, Via delle Pandette 32 - 50127 Firenze, Italy, paolo.brunori@unifi.it.

‡ifo Munich and LMU Munich, hufe@ifo.de.

§University of Copenhagen and World Bank, dmahler@worldbank.org.

1 Introduction

John Roemer's (1998) seminal contribution, *Equality of Opportunity*, has incited a flourishing empirical literature on the measurement of unequal opportunities. At the heart of Roemer's formulation is the idea that factors that determine individual outcomes can be divided into two categories: factors over which individuals have control, which he calls *effort*, and factors for which individuals cannot be held responsible, which he calls *circumstances*. Individuals characterized by exactly the same exogenous circumstances are assigned to a circumstance *type*. Members of a type have the same background conditions to transform resources into outcomes. Therefore, while within-type inequality, as caused by the differential exertion of effort, is morally irrelevant, between-type differences in achievements are inequitable and call for compensation. Thus, opportunity-equalizing policies have the objective of neutralizing the impact of circumstances on the distribution of the desirable outcome.

Following Roemer's approach, a battery of methods to measure inequality of opportunity have been proposed (see Roemer and Trannoy, 2015; Van de gaer and Ramos, 2016, for recent overviews).[1] Today, well established empirical methods include summary indexes that quantify the extent of unequal opportunities (Almås et al., 2011; Bourguignon et al., 2007; Checchi and Peragine, 2010) as well as statistical tests that detect the mere existence thereof (Kanbur and Snell, 2017; Lefranc et al., 2009). In either case, empirical results are sensitive to critical choices of model selection which are under complete discretion of the researcher.

First, researchers have to make a decision on which circumstance variables to consider for estimation.[2] Observable circumstances beyond individual control are typically a subset of the real number of exogenous variables affecting individual outcomes. This issue has been largely discussed by the literature, and the prevailing view is that partial observability implies downward-biased inequality of opportunity estimates (Ferreira and Gignoux, 2011). To counteract this downward bias, one strategy is to resort to high-quality datasets that provide very detailed information with respect to individual circumstances (Hufe et al., 2017). Naturally, the scope of improvement of this approach is limited by sample sizes. Consider for example the increasing availability of genetic datasets with billions of polymorphisms per person (Altshuler et al., 2015). While the genetic make-up of individuals clearly is beyond individual control and must be considered a circumstance, the number of circumstances exceeds the available degrees of freedom which forces the researcher to choose selectively from the available set of circumstances.

Second, the influence of circumstances may be dependent on the expression of other circumstance characteristics. 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 (García et al., 2017). In contrast to such evidence, however, many empirical applications presume that the effect of circumstances on individual outcome is fixed and additive (Bourguignon et al., 2007; Ferreira and Gignoux, 2011). On the one hand, analogous to partial observability, this functional form assumption forces a downward bias on inequality of opportunity estimates. On the other hand, limitations in the available degrees of freedom may prove the estimation of fully saturated models impractical. Again the researcher is left to her own devices in selecting the best model for estimating inequality of opportunity.

While the downward bias of inequality of opportunity estimates is prominently discussed in the extant literature, the reliability of estimates has been largely disregarded. Holding the

[1] Note that a number of contributions from the social choice literature on fair allocation had previously proposed similar methods (Fleurbaey, 1995, 2008; Van de gaer, 1993).

[2] Roemer does not provide a fixed list of variables that are to be considered as circumstances. Rather he suggests that the set of circumstances should evolve from a political process (Roemer and Trannoy, 2015). In empirical implementations typical circumstances are biological sex, socioeconomic background, race, or the area of birth.

sample size constant, increasing the type partition by including additional circumstances or relaxing the linearity assumption directly translates into reduced variation for estimating the relevant parameters. In fact, Brunori et al. (2016) show that overfitting the model may instill an upward bias on inequality of opportunity estimates.

This discussion highlights the non-trivial challenge in selecting the appropriate model for estimating inequality of opportunity. Scholars must balance between different sources of bias while trying to avoid ad-hoc solutions.

In this paper we propose the use of classification and regression tree methods to address the outlined shortcomings of current approaches. Introduced by Morgan and Sonquist (1963) and popularized by Breiman et al. (1984), classification and regression trees belong to a family of statistical methods that are commonly summarized under the labels of "machine learning" or "statistical learning" (Friedman et al., 2009). Originating from the fields of computer science and statistics, these methods are being increasingly adopted by economists (Athey, 2017; Mullainathan and Spiess, 2017; Varian, 2014). Classification and regression trees methods were developed to make out-of-sample predictions of a dependent variable based on a number of observable predictors. They let algorithms automatically segment the predictor space into non-overlapping regions to find the best model for predicting the outcome of interest. In the context of estimating equality of opportunity, this means that we let an automated algorithm decide how to partition the population into mutually exclusive types for the purpose of calculating measures of inequality of opportunity in the spirit of Roemer's theory. To be precise, within the class of classification and regression tree methods we focus on *conditional inference trees* and *conditional inference forests*, both of which bear a number of substantial advantages (Hothorn et al., 2006).

First, by drawing on a clear-cut algorithm one minimizes the degree of arbitrariness in model selection. In both trees and forests types are obtained in the attempt to explain outcome variability without assuming anything about which circumstances play a statistically significant role in shaping individual opportunities and how they interact. Thus, the partition of the population into Roemerian types is no longer a judgment call of the researcher but a non-arbitrary outcome of data analysis. Second, the conditional inference algorithm branches trees (and constructs forests) by using a sequence of hypothesis tests that prevent model overfitting. Therefore, by using the conditional inference algorithm we can both derive a test for the null hypothesis of equal opportunity and avoid the potential upward bias of inequality of opportunity estimates. As a consequence of avoiding upwards and downwards biases, and in contrast to the current approaches, our estimates are better suited for comparisons across time and between countries when sample sizes differ.

Aside from those shared merits, trees and forests bear some distinct advantages which the researcher needs to trade off when selecting her preferred approach within the class of conditional inference methods. Trees, on the one hand, have intuitive appeal and their graphical illustrations are instructive tools for longitudinal or cross-sectional comparisons of opportunity structures. Forests, on the other hand, perform better in trading off the different sources of bias outlined above. In fact, we will show that conditional inference forests outperform other prevalent estimation techniques in terms of out-of-sample prediction accuracy.

To be sure, just as the literature on intergenerational mobility (Black and Devereux, 2011), scholars of equality of opportunity are reluctant to give their estimates a causal interpretation. The ambition of the literature is to understand how much variation in outcomes can be attributed to root causes that are commonly perceived as unfair. It is precisely the *prediction* character of these empirical exercises that makes this branch of the literature a useful field to leverage the advantages of machine learning algorithms.

The remainder of this paper is organized as follows: section 2 gives a brief introduction with respect to current empirical approaches in the literature. Section 3 introduces regression trees

and illustrates how to use 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 4, in which we will also highlight the particular advantages of using tree-based estimation methods. Lastly, section 5 concludes.

2 Empirical Approaches to Equality of Opportunity

Consider a population of size N indexed by $i \in \{1, ..., N\}$ and an associated vector of incomes $Y = \{y_1, ..., y_i, ..., y_N\}$. Individual i's outcome of interest y_i is the result of two sets of factors. First, a set of *circumstances* beyond her control of size P: $\Omega_i = \{C_i^1, ..., C_i^p, ..., C_i^P\}$. Second, a set of *efforts* of size Q: $\Theta_i = \{E_i^1, ..., E_i^q, ..., E_i^Q\}$. In general, the outcome generating function $g : \Omega \times \Theta \to \mathbb{R}_+$ can therefore be written as

$$y_i = g(\Omega_i, \Theta_i). \tag{1}$$

Each circumstance $C^p \in \Omega$ is characterized by a total of X^p possible realizations, where each realization is denoted as x^p. Based on the realizations x^p we can partition the population into a set of non-overlapping *types* $T = \{t_1, ..., t_m, ..., t_M\}$. A type is a subgroup of the original population uniform in terms of circumstances, i.e. individuals i and j belong to the same type $t_m \in T$ if $x_i^p = x_j^p \ \forall C^p \in \Omega$. They belong to different types $t_m \in T$ if $\exists C^p \in \Omega : x_i^p \neq x_j^p$. The number of types in the population is given by $M = \prod_{p=1}^{P} X^p$. Following Roemer (1998) we assume that the joint realizations of the effort variables $E^q \in \Theta$ can be summarized by a scalar $\pi \in [0, 1]$. Individuals sharing the same expression of effort are called a *tranche*. Hence, types and tranches define two particular ways of partitioning the population into subgroups, where group membership either indicates uniformity in circumstances (types) or effort (tranches).

In the literature we can distinguish two broad classes of equality of opportunity definitions.[3] First, the *ex-ante* view focuses on between-type differences in the value of opportunity sets without paying attention to the specific effort realizations of individual type members. According to this perspective, equality of opportunity is satisfied if the aggregate value of opportunity sets is equalized across types. One example in case is the *ex-ante utilitarian* perspective according to which the value of opportunity sets is indicated by the average outcome within the specific type. Thus, equality of opportunity would be realized if the mean outcome of each type was equal to the population mean. Second, the *ex-post* view focuses on individual outcomes conditional on effort exertion. According to this perspective, equality of opportunity would be satisfied if individual outcomes were equalized within each tranche, i.e. individuals with equal levels of effort exertion realize the same outcomes. A comprehensive discussion of the ex-ante and ex-post principles of equality of opportunity can be found in Fleurbaey and Peragine (2013). In the context of this paper we will restrict ourselves to the ex-ante utilitarian approach only.

Tests and Measures The extant literature has witnessed the development of empirical *tests* and *measures* for ex-ante utilitarian inequality of opportunity. A prominent example for the former category is provided by Lefranc et al. (2009), who show that rejecting the null hypothesis of no first-order stochastic dominance in type-specific outcome distributions is sufficient to reject the existence of equal opportunities in the population from an ex-ante utilitarian perspective. Furthermore, in a recent contribution Kanbur and Snell (2017) develop likelihood ratio tests that can serve to test for ex-ante utilitarian equality of opportunity. A widely adopted example of the latter category, is the measure developed by Van de gaer (1993) and Chec-

[3] Measures different from the ones illustrated here, have been proposed in the literature. The interested reader is referred to Van de gaer and Ramos (2016) for a comprehensive overview.

chi and Peragine (2010). They propose to measure inequality in a counterfactual distribution $Y^{EA} = \{y_1^{EA}, ..., y_i^{EA}, ..., y_N^{EA}\}$ obtained by removing inequality within types from the original distribution. To be precise, individual outcomes are re-scaled to match their respective type mean:

$$y_i^{EA} = \frac{1}{N_m} \sum_{i \in t_m} y_i = \mu_m, \ \forall i \in t_m, \forall t_m \in T, \tag{2}$$

where N_m is the size and μ_m the average outcome of type t_m. Therefore, any remaining inequality in Y^{EA} reflects inequality between types and inequality of opportunity can now be summarized by applying any standard scalar measure of inequality $\mathrm{I}(\cdot)$, like the Gini index or a member of the generalized entropy class (Cowell, 2016), to the counterfactual distribution Y^{EA}. Any such measure obtains its minimal value in the case of equality of all type means, i.e if $\mu_m = \mu_l = \mu \ \forall t_m, t_l \in T$.

Estimation In practice we do not observe the full set of circumstances Ω. Rather we observe the subset $\check{\Omega} \subseteq \Omega$ of size $\check{P}$. For example, in most datasets we do not have full information on the genetic make-up of individuals neither do we have a gapless documentation of the socioeconomic conditions in which individuals grew up. Analogously, for most $C^p \in \check{\Omega}$ we only observe the subset $\check{X}^p$ of the true number of realizations X^p. For example, in many datasets information on parental education and occupation is coded in categorical variables of varying detail, which may mask more nuanced socioeconomic differences among households.

Depending on the strength of their distributional assumptions, estimations of inequality of opportunity are typically classified as either non-parametric or parametric. A point in case for the former approach is the abovementioned measure put forward by Van de gaer (1993) and Checchi and Peragine (2010). The researcher partitions the sample into mutually exclusive cells based on the realizations of all circumstance variables under consideration. Hence, the researcher makes no assumption on the interaction of circumstance variables in the determination of individual outcomes. This comes at a high cost, however. To avert overfitting, the partition must be constructed such that a sufficient number of observations belongs to each cell. Conditional on the dataset being rich enough in information on circumstances, this in turn forces the researcher to make a discretionary choice on the *relevant* partition. Consider for instance a continuous circumstance variable like parental income. Employing the non-parametric estimation approach, the researcher must split the parental income distribution into quantiles for constructing the type partition. The potential granularity of this split obviously depends on the sample sizes of the ensuing cells. Additionally, the researcher must balance the informational content of a finer partition of parental income against the opportunity cost of being forced to exclude another circumstance variable from the investigation. To put it in formal terms: the researcher must select a subset $\hat{\Omega} \subseteq \check{\Omega} \subseteq \Omega$ from the set of observed circumstances. Furthermore, within the confines of limited degrees of freedom the researcher must also decide for each $C^p \in \hat{\Omega}$, how to restrict the number of realizations $\hat{X}^p \subseteq \check{X}^p \subseteq X^p$ in order to construct a statistically meaningful type partition.

To address this problem, the literature commonly resorts to parametric estimation approaches. Here, the researcher obtains the counterfactual distribution by estimating a Mincerian regression with circumstances as the sole right-hand side variables (Bourguignon et al., 2007; Ferreira and Gignoux, 2011):

$$\ln(y_i) = \beta_0 + \sum_{p=1}^{\check{P}} \beta_p C_i^p + \epsilon_i. \tag{3}$$

The counterfactual distribution, Y^{EA} can then be constructed from the predicted values

$$y_i^{EA} = \exp\left[\sum_{p=1}^{\check{P}} \hat{\beta}_p C_i^p\right]. \tag{4}$$

Although the parametric approach solves some of the shortcomings of the non-parametric approach, it is not a panacea. The standard version of the parametric approach assumes a linear impact of all circumstances and therefore neglects the existence of interdependencies and non-linearities in the impact of circumstances. To pick up the example from the introduction, the researcher cannot allow for a differential impact of the same child-rearing arrangement on male and female children. Of course, to alleviate this shortcoming the researcher may integrate interaction terms and higher order polynomials into equation (3). At the extreme the researcher may even estimate a fully saturated model, in which case parametric and non-parametric estimation coincide. This congruence, however, reiterates the fundamental problem of current approaches towards the estimation of inequality of opportunity. In view of restrictions on the available degrees of freedom, the researcher is forced to make a discretionary choice on the model she estimates, which in itself is a strong determinant of the ensuing results when testing and measuring equality of opportunity. Furthermore, just as the non-parametric approach, the parametric estimation is at risk of overfitting the data when the set of circumstances is large.

In analogy to this paper, Li Donni et al. (2015) have discussed the issue of ad-hoc model selections in the empirical literature on equality of opportunity. To resolve this issue, they propose a data-driven type partition by estimating a latent class model. In this approach, observable circumstances are considered indicators of membership in an unobservable latent type, t_m. For each possible number of latent types, M, the model obtains the partition into types by minimizing the within-type correlation of observable circumstances, $C^p \in \check{\Omega}$. The optimal number of groups 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 these cases the latent class approach replicates the limitations of other prevalent approaches towards estimating inequality of opportunity: the researcher must pre-select the relevant set of circumstances, their subpartition as well as the respective interactions. To the contrary, our approach embodies a method to select circumstances from the set of all observed variables in a non-arbitrary fashion. Furthermore, latent types are constructed in the attempt to explain circumstances' correlation. The partition is therefore insensitive to the degree of association between circumstances and outcome. However, one may consider explaining outcome variability as function of circumstances to be precisely the purpose of inequality of opportunity measurement.[4] Lastly, we prefer the conditional inference approach as it provides the particular advantages of being econometrically more tractable while providing a stronger economic meaning of the identified types.

3 Estimating Inequality of Opportunity from Regression Trees

Originally, tree-based methods 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 equality of opportunity estimations and solve many of the

[4]This issue is common to any two-stage analysis in which latent classes serve as controls for a distal outcome. The effect of latent class membership on the distal outcome is attenuated and the explained variability is reduced (Lanza et al., 2013).

issues associated with the prevalent estimations approaches outlined in section 2. While we put a particular emphasis on *regression* trees, our main arguments also hold for *classification* trees. Thus, the proposed estimation methods are not restricted to continuous variables like income, but can also be fruitfully employed with respect to non-continuous outcomes, such as health (Trannoy et al., 2010) or education (Oppedisano and Turati, 2015).

In what follows we will present two tree-based estimation procedures both of which solve the model selection problem outlined in section 2. First, we will introduce conditional inference regression trees. Their simple graphical illustration is particularly instructive for longitudinal or cross-sectional comparisons of opportunity structures. In spite of their intuitive appeal, however, they perform relatively poorly in out-of-sample predictions. Second, to address the concern of estimate reliability we will also introduce conditional inference forests, which are – loosely speaking – a collection of many conditional inference trees. Forests do not have the intuitive appeal of regression trees. However, they perform significantly better in terms of out-of-sample predictions. In fact, we will show in section 4.5 that they outperform all other considered estimation techniques along this dimension.

3.1 Conditional Inference Trees

Tree-based methods obtain predictions for outcome y as a function of the input variables $I = \{I^1, ..., I^p, ...I^P\}$. Specifically, they use the set I to partition the population into a set of non-overlapping groups, $G = \{g_1, ..., g_m, ..., g_M\}$, where each group g_m is homogeneous in the expression of each input variable. These groups are also called *terminal nodes* or *leafs* in a regression tree context. The predicted value for outcome y of observation i is calculated from the mean outcome μ_m of the group g_m to which the individual is assigned. Hence, in addition to the observed income vector $Y = \{y_1, ..., y_i, ..., y_N\}$ one obtains a vector of predicted values $\hat{Y} = \{\hat{y}_1, ..., \hat{y}_i, ..., \hat{y}_N\}$, where

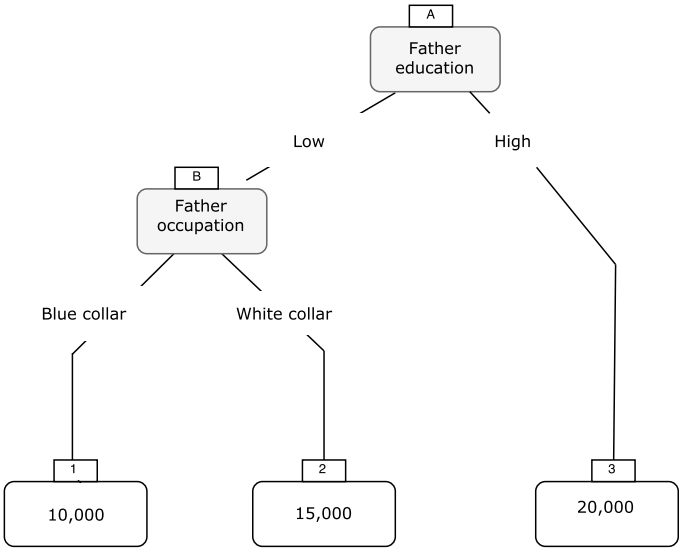
$$\hat{y}_i = \mu_m = \frac{1}{N_m} \sum_{i \in g_m} y_i, \ \forall i \in g_m, \forall g_m \in G. \tag{5}$$

The mapping from regression trees to equality of opportunity estimation is straightforward. Conditional on the input variables being circumstances only, i.e. $I \subseteq \check{\Omega} \subseteq \Omega$, it is evident that each resulting group $g_m \in G$ can be interpreted as a circumstance type $t_m \in T$. Furthermore, $\hat{Y}$ is analogous to the smoothed distribution Y^{EA}, the construction of which we have outlined in section 2 to illustrate ex-ante utilitarian measures of inequality of opportunity. In view of the fact that our predictor space is confined to circumstance variables only, we use the terms "input variables" and "circumstances" as well as "groups" and "types" interchangeably in the following. Input variables will be denoted by C^p and groups by t_m. In line with equation (5), we will refer to individual predictions $\hat{y}_i$ as μ_m.

Algorithm Considering all possible ways in which the population can be split into groups is a daunting task when the set of input variables is large. In conventional estimation approaches the researcher is left to her own devices in (i) selecting $\hat{\Omega}$ from $\check{\Omega}$, (ii) to restrict the number of realizations of each $C^p \in \hat{\Omega}$, and (iii) to determine the relevant interactions among all $C^p \in \hat{\Omega}$. The magnitude of this choice set oftentimes leads to arbitrary model selection. To the contrary, with regression trees the researcher does not need to make these choices herself. The researcher only submits the full and unrestricted set of observed variables that qualify as circumstances, $\check{\Omega}$, while the algorithm chooses the relevant circumstances, their subpartition and the respective interactions. To be precise, the observations are divided into M groups (or types) by what is known as *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 C^p. If C^p is a continuous or ordered variable, then $i \in t_m$ if $C_i^p < x^p$ and $i \in t_l$ if $C_i^p \geq x^p$. If C^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 $C^p \in \check{\Omega}$), 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 grey boxes indicate splitting points, while the white boxes indicate terminal nodes. The values inside the white boxes show predicted values associated with each terminal node (μ_m).

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 methodology proposed by Hothorn et al. (2006), leading to what they call conditional inference trees.[5]

Conditional inference trees are grown by a series of permutation tests according to the following 4-step procedure:

1. Test the null hypothesis of independence, $H_0^{C^p} : D(Y|C^p) = D(Y)$, for each input variable $C^p \in \check{\Omega}$, and obtain a p-value associated with each test, p^{C^p}.

   $\Rightarrow$ Adjust the p-values for multiple hypothesis testing, such that $p_{adj.}^{C^p} = 1 - (1 - p^{C^p})^P$ (Bonferroni Correction).

2. Select the variable, C^*, with the lowest adjusted p-value, i.e. $C^* = \{C^p : \text{argmin}\, p_{adj.}^{C^p}\}$.

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

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

[5] An alternative would be Classification and Regression Trees (CART) as introduced by Breiman et al. (1984). CART chooses splits so as to minimize the mean squared error, MSE $= \frac{1}{N} \sum_m \sum_{i \in t_m} (y_i - \mu_m)^2$. We prefer conditional inference trees since CART are biased towards splitting variables made of many categories (Hothorn et al., 2006). Furthermore, we avoid the intricacies of tree *pruning* (Friedman et al., 2009) by establishing a test criterion that considers the bias-variance trade-off before making an additional split.

3. Test the discrepancy between the subsamples for each possible binary partition, s, based on C^*, i.e. $Y_s = \{Y_i : C_i^* < x^p\}$ and $Y_{-s} = \{Y_i : C_i^* \geq x^p\}$, and obtain a p-value associated with each test, $p^{C_s^*}$.

   $\Rightarrow$ Split the sample based on $C_{s^*}^*$, by choosing the split point s that yields the lowest p-value, i.e. $C_{s^*}^* = \{C_s^* : \text{argmin}\, p^{C_s^*}\}$.

4. Repeat the algorithm for each of the resulting subsamples.

Conditional inference trees offer a particularly relevant structure in the context of inequality of opportunity. Each hypothesis test is essentially a test for whether equal opportunities exist within a particular (sub)sample. If the algorithm results in no splits at all, then we cannot reject the null hypothesis of equality of opportunity. The deeper the tree is grown, the more types are necessary to fully account for the inherent inequality of opportunities in the society under consideration. Each split tells us that the resulting types have significantly different opportunities under an ex-ante utilitarian interpretation. In all of the resulting types (i.e. the terminal nodes of the tree), we cannot reject the null of equal opportunities.

Tuning Note that the structure and depth of the resulting opportunity tree hinges crucially on the level of α. The less stringent the α-requirement, the more we allow for false positives, i.e. the more splits will be detected as significant and the deeper the tree will be grown. So how should α be chosen? On the one hand, α can be chosen a priori in line with the disciplinary convention to require significance levels of at least 5% or even 1%. On the other hand, we can let the data speak on the optimal specification of α, i.e. we can *tune* the α-parameter to find a model that performs optimally according to a pre-specified testing criterion.

If opting for the latter option, α is typically chosen by K-fold cross-validation (CV). To perform cross validation, one starts by splitting the sample into K subsamples, also called 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 real data points observed in the kth fold. The mean squared prediction error serves as an evaluation criterion:

$$\text{MSE}_k^{CV}(\alpha) = \sum_m \frac{N_m^k}{N^k} \sum_{i \in t_m} \frac{1}{N_m^k} (y_i^k - \mu_m(\alpha))^2. \tag{6}$$

This exercise is repeated for all K folds, so that $\text{MSE}^{CV}(\alpha) = \frac{1}{K} \sum_k \text{MSE}_k^{CV}(\alpha)$. One then chooses the α^* that delivers the lowest $\text{MSE}^{CV}(\alpha)$:

$$\alpha^* = \{\alpha \in A : \text{argmin } \text{MSE}^{CV}(\alpha)\}.^{6} \tag{7}$$

In our empirical application we fix $\alpha^* = 0.01$, which is in line with the disciplinary convention for hypothesis tests. However, we provide a sensitivity check using cross-validated α in Figure A.1 of Appendix A.3.

[6] One may argue that a criterion that evaluates models according to their capacity to predict *individual* outcomes is misplaced for ex-ante utilitarian inequality of opportunity estimations. Afterall, we have demonstrated above that we are mainly concerned with estimating type means rather than individual outcomes. In Appendix A.1 we show that the MSE-criterion and its handling of the variance-bias trade-off straightforwardly extends to balancing upward and downward biases in inequality of opportunity estimations.

3.2 Conditional Inference Forests

Regression trees solve the model selection problem outlined in section 2 and provide a simple and non-arbitrary way of dividing the population into types. Furthermore, trees are easily mapped and thus lay bare the opportunity structure of a given society for a larger audience. However, constructing the counterfactual distribution Y^{EA} from conditional inference trees suffers from two shortcomings: first, they only make limited use of the information inherent in the set of observed circumstances since not all $C^p \in \check{\Omega}$ are used for the construction of the tree. Yet, the omitted circumstances may possess some informational content that can increase predictive power even though they are not significantly associated with Y at level α^*. This is a particular issue if two important circumstances are highly correlated. Once a split is done using either of the two, the other will unlikely yield enough information to cause another split. Second, the predictions and thus the values of opportunity sets, μ_m, emanating from trees have a high variance. The structure of trees - and therefore the ensuing distribution Y^{EA} - is fairly sensitive to alternations in the respective data samples. This is a particular issue if there are various circumstances that are close competitors for defining the first split (Friedman et al., 2009). In what follows we will introduce conditional inference forests. Conditional inference forests build methodologically on conditional inference tress and are able to deal with both of these shortcomings (Breiman, 2001).

Algorithm In short, random forests create many trees and average over all of these when making predictions. Trees are constructed according to the same 4-step procedure outlined in the previous subsection. However, two tweaks are made. First, each tree is estimated on a random subsample b of the original data.[7] In total B such trees are estimated. Second, a random subset of circumstances $\bar{\Omega} \subseteq \check{\Omega}$ 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. Drawing only on subsets $\bar{\Omega} \subseteq \check{\Omega}$ increases the likelihood that all circumstances with informational content at some point will be identified as the splitting variable C^* and thus addresses the limited information use of regression trees. Furthermore, averaging over the B predictions cushions the variance of individual predictions μ_m and thus addresses the second shortcoming identified in relation with single regression trees. Therefore, predictions are formed as follows:

$$\hat{y}_i(\alpha, \bar{P}, B) = \frac{1}{B} \sum_{b=1}^{B} \mu_m^b(\alpha, \bar{P}). \tag{8}$$

Tuning From equation (8) it is evident 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. Again, these parameters can be imposed a priori by the researcher or they can be determined by tuning the three-dimensional grid $(\alpha, \bar{P}, B)$ 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.[8]

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

[7] Alternatively, one can draw bootstrapped samples, i.e. sample with replacement until a dataset with the same size as the original data is reached. We use the subsampling technique since it has been shown that using bootstrapping leads to biased variable selection (Strobl et al., 2007).

[8] Empirical tests show that this is the case with $B^* = 200$ for most countries in our sample (see Figure A.2 of Appendix A.3).

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 value for the hypothesis tests.

2. Calculate the average predicted value of observation i using each of the subsamples b_{-i} (the so called *bags*) in which i does not enter: $\hat{y}_i^{OOB}(\alpha, \bar{P}) = \frac{1}{B_{-i}} \sum_{b_{-i}} \mu_m^b(\alpha, \bar{P})$.

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

4. Choose $(\alpha^*, \bar{P}^*) = \{(\{\alpha \in A\}, \{\bar{P} \in \check{P}\}) : \text{argmin}\, \text{MSE}^{OOB}\}$.

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 b_{-i} that excludes all observations i. Hence, for each tree we can use the out-of-bag data points to evaluate the predictive accuracy of the respective model.[9]

The improved predictive quality of random forests comes at a cost. It is no longer possible to identify a fixed set of types T into which we can partition the population. For example, depending on the subset $\bar{\Omega} \subseteq \check{\Omega}$ used for a particular tree as well as the the particular subsample b drawn from the data, it may be that $i, j \in t_m^b$ but $i \in t_m^{b+1}$ while $j \notin t_m^{b+1}$. As a consequence, the individual prediction and hence the valuation of the individual opportunity set is an average over the value of opportunity sets μ_m^b associated with each tree of the forest. Therefore, the valuation of opportunity sets is less straightforward and opportunity structures are hard to illustrate in a graphical manner. It is nevertheless possible to describe opportunity structures by calculating the relative variable importance of each $C^p \in \check{\Omega}$ in constructing the forest. See section 4.3 for an illustration.

4 Empirical Application

In this section we provide an illustration of our methodology using harmonized survey data from 31 European countries. As outlined above, conditional inference trees and random forests solve the issue of model selection associated with the prevalent approaches to equality of opportunity estimations. Conditional inference trees are easily tractable and lend themselves to cross-sectional (and longitudinal) comparisons of opportunity structures. Conditional inference forests are less tractable but outperform the former approach in terms of predictive accuracy. In the following, we will illustrate the merits of both approaches. Furthermore, we will compare the results from both versions of our method with prevalent measurement approaches in the extant literature; namely parametric, non-parametric and latent class models. Comparisons will be made along two dimensions. First, the estimates themselves, and second, the respective out-of-sample accuracy. The latter criterion should be interpreted as an indicator of how well the respective method balances upward and downward biases in inequality of opportunity estimations. A formal argument for why this is the case, is provided in Appendix A.1.

4.1 Data

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

[9] In principle tuning can be conducted analogously to regression trees by means of k-fold cross validation. This, however, is computationally expensive. Cross-validation would require to repeat the entire estimation exercise for a total of K folds. This is not necessary when using the out-of-bag error since out-of-sample points are already delivered by leaving out observations i when using bag b_{-i}. Hence, in the case of forests using the out-of-bag error is K times more computationally efficient than cross-validation.

incomes, poverty, and living conditions on an annual basis and covers a cross-section of 31 European countries in the 2011 wave.[10] We draw on the 2011 wave since it contains an ad-hoc module about the intergenerational transmission of (dis)advantages, which allows us to construct finely-grained type partitions. The set of observed circumstances $\check{\Omega}$ and their respective expressions x^p are listed in Table 1 whereas descriptive statistics concerning circumstances are reported in Appendix A.2. As an additional advantage, EU-SILC has been extensively studied by the empirical literature on inequality of opportunity and thus provides appropriate benchmarks against which we can compare our method (Checchi et al., 2016; Marrero and Rodríguez, 2012; Palomino et al., 2016).

The unit of observation is the individual, whereas the outcome of interest is equivalent disposable household income. 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 our country sample. While households in Norway (NO) and Switzerland (CH) on average obtained incomes above €40,000 in 2010, the average households 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 are characterized by Gini coefficients of below 0.22. At the other end of the spectrum we find the Eastern European countries of Latvia (LV), Lithuania (LT) and Romania (EL) with Gini coefficients well above the level of 0.33.

4.2 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 (3), estimates are obtained by a Mincerian regression of equivalent household income on the following controls: father occupation (10 categories), father and mother education (five categories), area of birth (three categories), and tenancy status of the household (two categories). The model specification therefore includes 20 binary variables and resembles the specification used in Palomino et al. (2016).[11]

Second, we draw on the non-parametric approach as proposed by Checchi and Peragine (2010). Non-parametric estimates are obtained by partitioning the sample into 40 types. Individuals in type t_m have parents of equivalent education (five categories), share their migration status (a binary variable whether the respondent is a first or second generation immigrant), and have fathers working in the same occupation. To minimize the frequency of sparsely populated

[10] 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).

[11] We have estimated the predicted outcomes both as the exponential of the predicted log outcome, $y_i^{EA} = \exp\left[\sum_{p=1}^{\check{P}} \hat{\beta}_p C_i^p\right]$, and by introducing, assuming a normally distributed error term, the correction $y_i^{EA} = \exp\left[\sum_{p=1}^{\check{P}} \hat{\beta}_p C_i^p + \sigma^2/2\right]$, where σ^2 is the estimated variance of the error term. We do not find any significant differences in the level of estimated inequality of opportunity when introducing the correction. This may explain why the need of such correction has never been explicitly discussed in previous contributions.

Table 1: List of Circumstances

- Respondent's sex:
  - Male
  - Female
- Respondent's country of birth:
  - Respondent's present country of residence
  - European country
  - Non-European country
- Presence of parents at home:
  - Both present
  - Only mother
  - Only father
  - Without parents
  - Lived in a private household without any parent
- Number of adults (aged 18 or more) in respondent's household
- Number of working adults (aged 18 or more) in respondent's household
- Number of children (under 18) in respondent's household
- Father/mother 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
- Father/mother education (based on 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)
- Father/mother occupational status:
  - Unknown or dead father/mother
  - Employed
  - Self employed
  - Unemployed
  - Retired
  - House worker
  - Other inactive
- Father/mother main occupation (based on 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)
  - Father/mother did not work, was unknown or was dead (I-0)
- Managerial position of the father/mother:
  - Supervisory
  - Non-supervisory
- Tenancy status of the house in which the respondent was living:
  - Owned
  - Not owned

types we divert from the occupational list given in Table 1 by re-coding occupations into three categories: highly skilled non-manual (I-01–I-03), lower-skilled non-manual (I-04–I-05 and I-10), skilled manual and elementary occupation (I-06–I-09 and father/mother unknown or dead). This

Table 2: Summary Statistics

Country	Sample size	Avg. eq. income	Std. dev.	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

Note: Summary statistics for the 31 countries in the 2011 wave of EU-SILC. Income variables are measured in Euros.

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.

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, $\check{\Omega}$. We follow Li Donni et al. (2015) in using Schwartz's Bayesian Information Criterion (BIC) to select the most adequate number of latent types.

4.3 Estimates of Inequality of Opportunity

Table 3 shows inequality of opportunity estimates for our country sample according to five different estimation procedures. Columns 2-4 list results using the parametric, the non-parametric, and the latent class approach, all of which have been proposed in the extant literature (see section 4.2). Columns 5 and 6 list results from conditional inference trees and conditional inference forests, respectively. For all methods, inequality of opportunity estimates are obtained by calculating the Gini index in the counterfactual distribution Y^{EA}.

Of all methods under consideration the parametric approach delivers the highest estimates. For 29 out of 31 countries the inequality of opportunity estimates are higher than the results

from both conditional inference trees and forests. Analogously, the unweighted average estimate over all countries equals 0.103 Gini points for the parametric approach as compared to 0.079 and 0.078 Gini points for trees and forests, respectively. Also in terms of country rankings, the parametric approach delivers markedly different results in comparison to our preferred methods. 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, 1st and 5th (6th, 2nd and 7th) in the case of trees (forests).

Non-parametric measures of inequality of opportunity take a middle-ground between the parametric approach and our preferred methods. For 16 (19) out of 31 countries the non-parametric estimate exceeds the estimate coming from trees (forests), while the unweighted cross-country average estimate amounts to 0.084 Gini points. In terms of country rankings the non-parametric approach shows much closer resemblance to our preferred methods than the parametric approach. For example, the three most unequal countries from an opportunity perspective as identified by the non-parametric approach are Bulgaria (BG), Portugal (PT) and Luxembourg (LU), which is congruent with the top three countries identified by trees and forests.

Lastly, the latent class model tends to furnish much lower estimates than all other methods, including trees and forests. This is not very surprising if one considers how latent types are constructed. Latent classes are obtained in the attempt to maximize local independence, that is to minimize the within-type correlation of circumstances. The algorithm constructs types (and selects their most appropriate number) ignoring the covariance of circumstances and outcome. Conditional inference trees instead construct types by maximizing the outcome variability that can be explained by circumstances. For 8 (9) out of 31 countries the latent class estimate falls short of the estimate coming from trees (forests), while the unweighted cross-country average estimate amounts to 0.069 Gini points. Also in terms of country rankings the latent class approach differs markedly from our preferred methods. It identifies Romania (RO), Greece (EL) and Portugal (PT) as the countries in which opportunities are most unequally distributed, whereas these countries rank 6th, 5th and 1st (6th, 7th and 3rd) in the case of trees (forests).

To gain further understanding as regards the relation of existing measurement approaches to our preferred methods, Figure 2 plots the estimates from each method against the estimates from conditional inference forests. The black diagonal indicates the 45 degree line, on which all data points should align if the different methods were perfectly congruent. The upper left panel plots the estimates from the parametric approach against the forest estimates. We can confirm the previous diagnosis that the parametric approach delivers higher estimates than forests (and trees). The difference is particularly pronounced for countries that are characterized by relatively low levels of inequality of opportunity, like the Nordic countries. The upper right panel shows the same plot for the non-parametric approach. We again find relatively high upward divergences in comparison to conditional forest estimates for countries in which inequality of opportunity is low. However, the differences are less pronounced. Interestingly, this pattern in reversed when looking at the correlation plot for the latent class approach in the lower left panel. Instead of overestimating the impact of circumstances in societies of low inequality of opportunity, it underestimates the impact of circumstances in societies that are characterized by high inequality in opportunities. Finally, as expected, trees and forests tend to produce very similar results. The correlation between estimates is high (0.98) and in contrast to all other approaches, the sign of the difference is uncorrelated with the level of the estimate.

4.4 Opportunity Structure

Endowed with an estimate of inequality of opportunity, adequate policy responses must be informed by the particular opportunity structure of a society. That is, policy makers want to

Table 3: Inequality of Opportunity Estimates

Country	Parametric	Non-Parametric	Latent Class	Cond. Inf. Tree	Cond. Inf. Forest
AT	0.0888	0.0751	0.0796	0.0865	0.0880
BE	0.1108	0.0868	0.0534	0.0868	0.0912
BG	0.1542	0.1356	0.1148	0.1362	0.1335
CH	0.0917	0.0827	0.0631	0.0796	0.0901
CY	0.0942	0.0831	0.0738	0.0799	0.0800
CZ	0.0716	0.0659	0.0600	0.0569	0.0511
DE	0.0704	0.0588	0.0467	0.0697	0.0793
DK	0.0772	0.0409	0.0289	0.0212	0.0204
EE	0.1108	0.1020	0.0744	0.0967	0.1005
EL	0.1476	0.1208	0.1165	0.1264	0.1089
ES	0.1421	0.1201	0.0893	0.1280	0.1200
FI	0.0687	0.0515	0.0475	0.0197	0.0275
FR	0.0858	0.0863	0.0717	0.0904	0.0980
HR	0.1312	0.0884	0.0758	0.0822	0.0763
HU	0.1098	0.1033	0.0951	0.1134	0.1079
IE	0.1048	0.0971	0.0484	0.0843	0.0784
IS	0.0669	0.0321	0.0297	0.0123	0.0157
IT	0.1213	0.0907	0.0799	0.1078	0.0969
LT	0.0947	0.0674	0.0587	0.0693	0.0672
LU	0.1340	0.1209	0.0904	0.1326	0.1356
LV	0.1335	0.1099	0.0951	0.1102	0.1110
MT	0.0872	0.0796	0.0566	0.0710	0.0716
NL	0.0661	0.0529	0.0411	0.0284	0.0194
NO	0.0480	0.0405	0.0296	0.0202	0.0234
PL	0.1111	0.0973	0.0953	0.1019	0.0991
PT	0.1376	0.1236	0.1156	0.1362	0.1267
RO	0.1698	0.1040	0.1194	0.1204	0.1107
SE	0.1178	0.0604	0.0251	0.0247	0.0313
SI	0.0772	0.0730	0.0588	0.0317	0.0361
SK	0.0626	0.0507	0.0420	0.0495	0.0459
UK	0.1012	0.0896	0.0622	0.0714	0.0791

Note: Estimates of inequality of opportunity using five different estimation methods. Inequality of opportunity is measured as the Gini coefficient in the counterfactual distribution Y^{EA}.

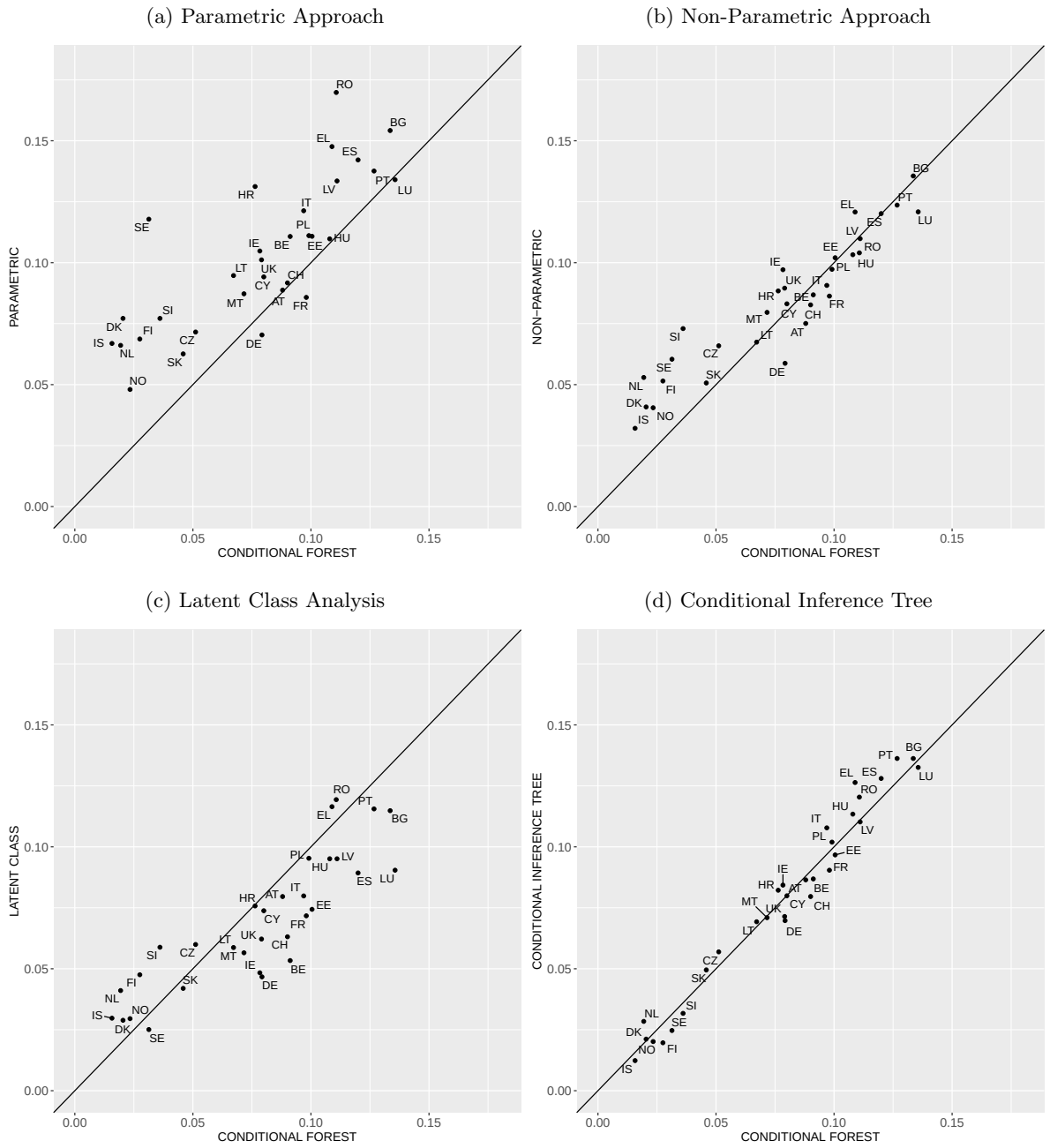
learn about the particular circumstance types the differences among which cause the existence of inequality of opportunity. In this section we illustrate such analyses for both trees and forests. To keep the analysis intelligible we restrict ourselves to two interesting cases: Sweden and Germany.

Trees As outlined in section 3.1, the analysis of opportunity structures is particularly intuitive in the case of trees as the relevant information can be directly read off their graphical illustration.

Figure 3 illustrates the opportunity structure of Sweden that 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 Europe and the collective group of native residents and European immigrants. The former type accounts for about 10% of the population and on average obtains an equivalent household income that is 35% lower than the corresponding income of the latter group. Recall that each split is based on a statistical test for the existence of equality of opportunity within the respective internal node. Thus, in Sweden we can reject the null hypothesis of equal opportunities for first-generation immigrants born outside Europe and the remainder of the population. However, within these sub-groups the null hypothesis of equality of opportunity cannot be rejected.

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

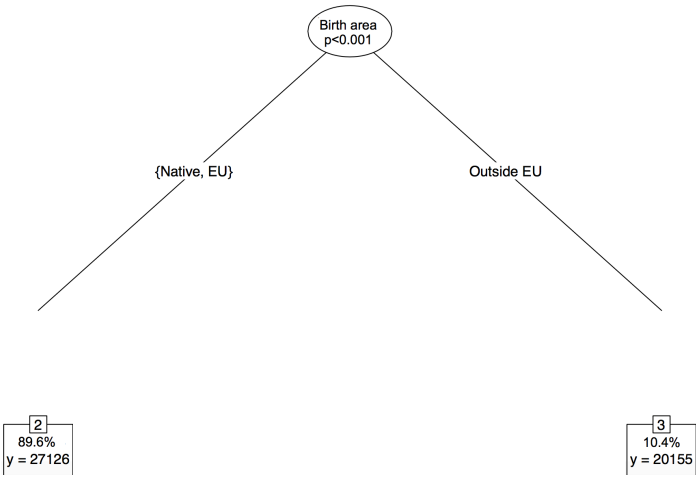
Figure 2: Correlation of Estimates by Method



Note: Comparison of inequality of opportunity estimates based on random forests with estimates based on four other methods. Along the solid line inequality of opportunity is the same for the two methods.

tion, 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.

Figure 3: Opportunity Tree: Sweden



Note: Opportunity tree for Sweden. White rectangular boxes indicate terminal nodes. The first number inside the rectangular boxes indicates the share of the population belonging to this group, while the second number indicates the predicted income.

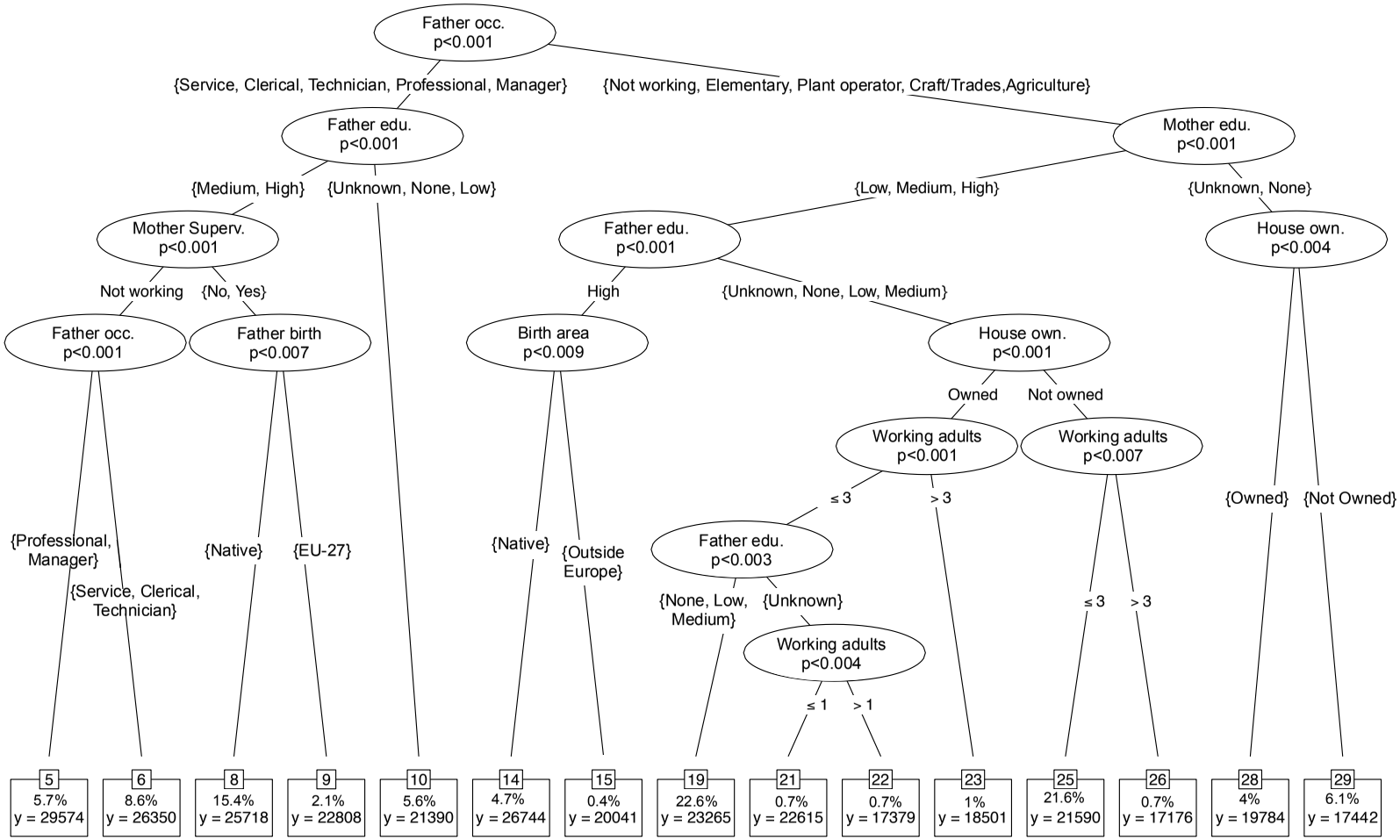
If fathers of respondents had no or low education, the offspring earned less (€21,390) than the country average in spite of the fact that fathers made a career in a high-rank occupation. Conditional on the father both being highly educated and working in a high-rank occupation, the intra-household division of labor plays a strong 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 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 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-0). 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 Europe (Terminal node 15). Overall, when analyzing the right-hand side of the tree, it is clear that circumstances interact in a very different way in determining individuals' outcomes. In addition to parental education and the migration status of individuals, the tenancy status during childhood as well as the number of working adults in the household play an important role.

There is marked heterogeneity in tree structures across countries. For the remaining countries in our sample, terminal nodes range from three (Denmark, Iceland and Norway) to 27 (Italy).[12] It is noteworthy that the rank-rank correlation between the number of terminal nodes and the inequality of opportunity estimates presented in section 4.3 is positive but not perfect (Appendix A.5).

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

[12]Figures of the tree structures for the remaining countries are available upon request.

Figure 4: Opportunity Tree: Germany

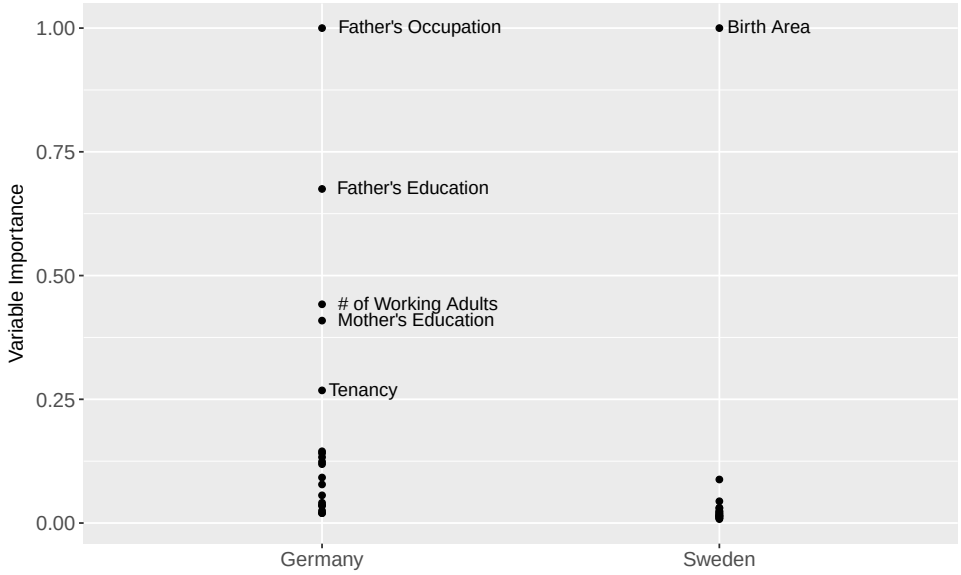


Note: Opportunity tree for Germany. White rectangular boxes indicate terminal nodes. The first number inside the rectangular boxes indicates the share of the population belonging to this group, while the second number indicates the predicted income. Occupation refers to ISCO-08 one digit codes. All variables describing household characteristics refer to the period in which the respondent was about 14 years old. See Table 1 for details.

construction of opportunity forests. One measure of variable importance, as proposed by Strobl et al. (2007), is obtained by permuting input variable $C^p \in \check{\Omega}$ 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 $C^p \in \check{\Omega}$ for prediction accuracy. Repeating this procedure for all $C^p \in \check{\Omega}$ affords a relative comparison of all circumstances.

Figure 5 shows the results from this procedure for our example cases of Germany and Sweden. Each black dot is the importance of one of the $\check{P}$ variables in the set of observed circumstances $\check{\Omega}$. 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.

Figure 5: Variable Importance for Germany and Sweden



Note: Each dot shows the importance of a particular circumstance for the predictions from our random forest. The importance of a circumstance is measured by permuting the circumstance, calculating a new MSE^{OOB}, and computing the difference in the MSE^{OOB} between the original model and the model with the permuted circumstance. The importance measure is standardized such that the circumstance with the greatest importance in each country equals one. Occupation refers to ISCO-08 one digit codes. All variables describing household characteristics refer to the period in which the respondent was about 14 years old. See Table 1 for details.

It is reassuring that these findings are in line with the graphical analysis of opportunity trees. In Figure A.3 of Appendix A.4 we show variable importance plots for all countries in our sample. 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 for which father's occupation and father's education are most important. Third, there is a group of Eastern European countries for which mother's education and occupation is most important.

4.5 Out-of-Sample Performance

Recall that current approaches towards estimating inequality of opportunity are subject to different biases. Models are downward biased to the extent that the full set of circumstances Ω is unobserved. Models are upward biased to the extent that they over-utilize the set of observed circumstances $\check{\Omega}$ leading to overfitted estimates that do not replicate out-of-sample (see Appendix A.1 for the formal argument).

In order to assess how well different estimation approaches trade off these biases, we follow the machine-learning practice of splitting our sample into a *training set* with $i_{train} \in \{1, ..., N_{train}\}$ and a *test set* with $i_{test} \in \{1, ..., N_{test}\}$. For each country in our sample, $N_{train} = \frac{2}{3}N$ while $N_{test} = \frac{1}{3}N$. 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 3.1 for trees, section 3.2 for forests, and section 4.2 for our benchmark methods).
2. Store the prediction functions $\hat{f}_{train}(\check{\Omega})$.
3. Predict the outcomes of observations in the test set: $\hat{y}_{i_{test}} = \hat{f}_{train}(\check{\Omega}_{i_{test}})$.
4. Calculate the out-of-sample error: $\text{MSE}^{test} = \frac{1}{N_{test}} \sum_{i_{test}} [y_{i_{test}} - \hat{y}_{i_{test}}]^2$.

Figure 6 compares the resulting MSE^{test} of the different models. For each country, the MSE^{test} of random forests is standardized to equal 1, such that an MSE^{test} larger than 1 represents a worse fit out-of-sample. That implies that the respective method performs worse than forests in trading off upward and downward biases, either by neglecting the use of circumstances or overfitting. We derive 95% confidence intervals based on 200 bootstrapped re-samples of the test data using the normal approximation method.

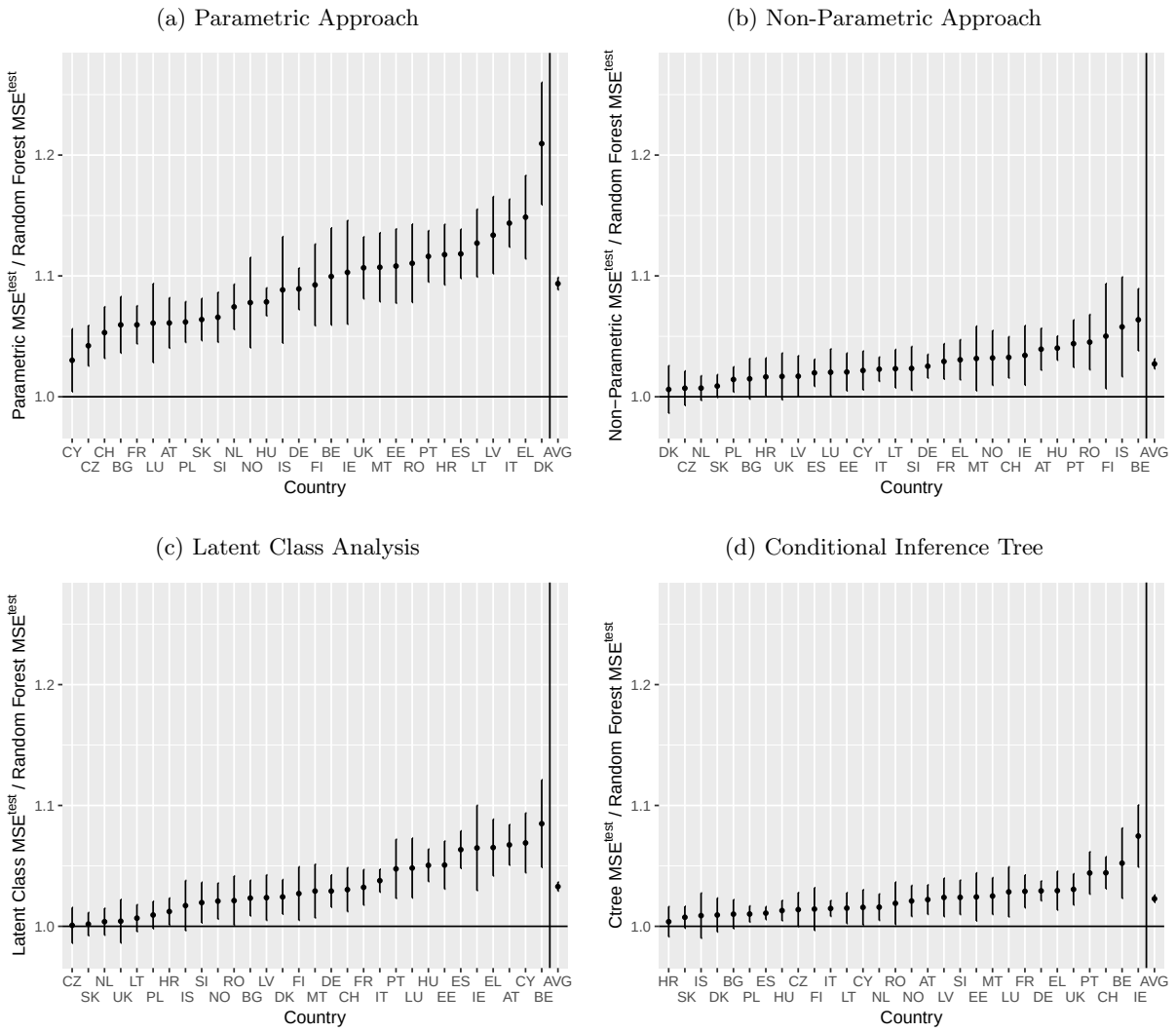
As expected, random forests outperform all other methods in nearly all cases. On average, the parametric approach gives a fit 9.4% worse than forests. With average shortfalls of around 3%, out-of-sample prediction errors are less pronounced for non-parametric models and latent class analysis. Yet both methods perform worse than conditional inference forests for the vast majority of countries in our sample. Hence, relative to random forests, our benchmark methods either underutilize or overutilize the information contained in $\check{\Omega}$ and are therefore biased in their inequality of opportunity estimates. The estimates presented in section 4.3 suggest that the parametric and the non-parametric partitions are overfitting the data, while the type partition delivered by latent class analysis is too coarse.

On average conditional inference trees are closest to the test error rate of forests. Yet they also fall short of the performance of forests due to their poorer utilization of the information given in $\check{\Omega}$.

5 Conclusion

In this paper we have proposed two novel approaches towards estimating inequality of opportunity based on regression trees. Both conditional inference trees and forests minimize arbitrary model selection by the researcher, while trading off downward and upward biases in inequality of opportunity estimates. On the one hand, 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. On the other hand, conditional inference trees are econometrically less complex and provide a handy graphical illustration that can be used for the straightforward analysis of opportunity structures. The fact that trees are very close to forests in terms of their inequality of opportunity estimates (section 4.3), the importance they assign to

Figure 6: Comparison of Models' Test Error



Notes: The figure compares the test error of the different models. The test error of random forests is standardized to 1, such that a test error larger than 1 represent worse fits than random forests. 95% confidence intervals are derived based on 200 bootstrapped re-samples of the test data using the normal approximation method. 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.

specific circumstances (4.4) and their out-of-sample performance (4.5) 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 we proposed 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. 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 Lefranc et al. (2009), provides an interesting avenue for future research.

References

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

Altshuler, D., Durbin, R. M., Donnelly, P., Green, E. D., Nickerson, D. A., Boerwinkle, E., and Doddapaneni, H. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Athey, S. (2017). The Impact of Machine Learning on Economics. *mimeo*.

Black, S. E. and Devereux, P. J. (2011). Recent Developments in Intergenerational Mobility. In Card, D. and Ashenfelter, O., editors, *Handbook on Labor Economics*, volume 4, chapter 16, pages 1487–1541. Elsevier, Amsterdam.

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

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

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

Brunori, P., Peragine, V., and Serlenga, L. (2016). Upward and downward bias when measuring inequality of opportunity. *ECINEQ Working Paper Series*, 2016-406.

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

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

Cowell (2016). Inequality and Poverty Measures. In Adler, M. D. and Fleurbaey, M., editors, *Oxford Handbook of Well-Being and Public Policy*, chapter 4, pages 82–125. Oxford University Press, Oxford.

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

Ferreira, F. H. G. and Gignoux, J. (2011). The Measurement of Inequality of Opportunity: Theory and an Application to Latin America. *Review of Income and Wealth*, 57(4):622–657.

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

Fleurbaey, M. (2008). *Fairness, Responsibility, and Welfare*. Oxford University Press, Oxford.

Fleurbaey, M. and Peragine, V. (2013). Ex Ante Versus Ex Post Equality of Opportunity. *Economica*, 80(317):118–130.

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

García, J. L., Heckman, J. J., and Ziff, A. L. (2017). Gender Differences in the Benefits of an Influential Early Childhood Program. *European Economic Review*, Forthcoming.

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

Hufe, P., Peichl, A., Roemer, J. E., and Ungerer, M. (2017). Inequality of income acquisition: the role of childhood circumstances. *Social Choice and Welfare*, Forthcoming.

Kanbur, R. and Snell, A. (2017). Inequality Measures as Tests of Fairness. *mimeo*.

Lanza, S., Xianming, T., and Bethany, B. (2013). Latent class analysis with distal outcomes: A flexible model- based approach. *Structural Equation Modeling*, 20(1):1–26.

Lefranc, A., Pistolesi, N., and Trannoy, A. (2009). Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France. *Journal of Public Economics*, 93(11–12):1189–1207.

Li Donni, P., Rodríguez, J. G., and Rosa Dias, P. (2015). Empirical definition of social types in the analysis of inequality of opportunity: A latent classes approach. *Social Choice and Welfare*, 44(3):673–701.

Marrero, G. A. and Rodríguez, J. G. (2012). Inequality of Opportunity in Europe. *Review of Income and Wealth*, 58(4):597–621.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, 58(302):415–434.

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

Oppedisano, V. and Turati, G. (2015). What are the causes of educational inequality and of its evolution over time in Europe? Evidence from PISA. *Education Economics*, 23(1):3–24.

Palomino, J. C., Marrero, G. A., and Rodríguez, J. G. (2016). Channels of inequality of opportunity: The role of education and occupation in Europe. *ECINEQ Working Paper Series*, 2016-411.

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

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

Roemer, J. E. and Trannoy, A. (2015). Equality of Opportunity. In Atkinson, A. B. and Bourguignon, F., editors, *Handbook of Income Distribution*, volume 2, chapter 4, pages 217–300. Elsevier, Amsterdam.

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

Trannoy, A., Tubeuf, S., Jusot, F., and Devaux, M. (2010). Inequality of opportunities in health in France: A first pass. *Health Economics*, 19(8):921–938.

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

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

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

A Appendix

A.1 Model Evaluation by the MSE

We use the MSE as a model evaluation criterion when cross-validating α in the case of trees (Section 3.1) and when determining the values of α and $\bar{P}$ using the out-of-bag error rate in the case of forests (Section 3.2). Analogously, when comparing the predictive performance of different estimation approaches in the test sample N^{test}, we prefer the estimation approach that yields a lower MSE (Section 4.5). The following discussion applies to all of these applications. For the sake of conciseness, superscript h always indicates observations in the hold-out sample regardless of the specific application.

The general MSE evaluation criterion can be written as follows:

$$\frac{1}{N^h}\sum_h (y_i^h - \hat{y}_i)^2. \tag{9}$$

In the case where observed circumstances $\check{\Omega}$ are the sole input variables, individual predictions $\hat{y}_i$ are given by the mean outcomes of the type to which individuals are allocated and we can write:

$$\frac{1}{N^h}\sum_h (y_i^h - \mu_m)^2, \tag{10}$$

where $\mu_m = \frac{1}{N}\sum_{i\in t_m} y_i$ and t_m denotes a specific type in the model we want to evaluate. It is instructive to rewrite the MSE as a weighted average over types as follows:

$$\sum_m \frac{N_m^h}{N^h}\sum_{i\in t_m}\frac{1}{N_m^h}(y_i^h - \mu_m)^2. \tag{11}$$

We can expand the previous expression and spell out the binomial formula:

$$\sum_m \frac{N_m^h}{N^h}\sum_{i\in t_m}\frac{1}{N_m^h}\left[(y_i^h - \mu_m^h) + (\mu_m^h - \mu_m)\right]^2 \tag{12}$$

$$= \sum_m \frac{N_m^h}{N^h}\sum_{i\in t_m}\frac{1}{N_m^h}\left[(y_i^h - \mu_m^h)^2 + (\mu_m^h - \mu_m)^2\right] + 2\sum_m \frac{N_m^h}{N^h}(\mu_m^h - \mu_m)\sum_{i\in t_m}\frac{1}{N_m^h}(y_i^h - \mu_m^h). \tag{13}$$

Evidently, $\sum_{i\in t_m}\frac{1}{N_m^h}(y_i^h - \mu_m^h) = 0$ and the formula simplifies to:

$$\sum_m \frac{N_m^h}{N^h}\sum_{i\in t_m}\frac{1}{N_m^h}\left[\underbrace{(y_i^h - \mu_m^h)^2}_{(1)} + \underbrace{(\mu_m^h - \mu_m)^2}_{(2)}\right], \tag{14}$$

where (1) is the intra-type variance of outcomes in the hold-out sample and (2) is the variance of type-means between the hold-out sample and the training sample. Recall that we prefer models that minimize formula (14). For the sake of exposition, let's generalize the previous expression by introducing the weighting parameter $\delta \in [0, 1]$. Note that the standard MSE yields equivalent rankings to the special case in which $\delta = 0.5$, i.e. the case in which we give equal weight to both (1) and (2):

$$\sum_m \frac{N_m^h}{N^h}\sum_{i\in t_m}\frac{1}{N_m^h}\left[\delta\underbrace{(y_i^h - \mu_m^h)^2}_{(1)} + (1-\delta)\underbrace{(\mu_m^h - \mu_m)^2}_{(2)}\right]. \tag{15}$$

Now assume two extreme cases:

1. $\delta = 1$: In this case we give full priority to (1), i.e. we would always prefer a model that decreased the intra-type variance in the hold-out sample the most. Naturally, one reduces intra-type variance by increasing the granularity of the type partition. Hence, we would always prefer the model that used more circumstances and interactions. Thus, (1) addresses the downward bias of equality of opportunity estimates as induced by using only a subset $\hat{\Omega}$ of the full set of observed circumstances $\check{\Omega}$.

2. $\delta = 0$: In this case we give full priority to (2), i.e. we would always prefer a model that decreased the variance between type means in the hold-out sample and the type means in the training sample. Invoking the law of large numbers it is evident that the ideal model from this perspective is the model with no partition at all, i.e. the model in which individual predictions μ_m are given by the sample mean μ. Thus, (2) addresses the upward bias identified by Brunori et al. (2016) that originates from overfitting the model to the training data.

To conclude, the more weight we put on (1), the less the downward bias in our estimation, since we allow circumstances to have explanatory scope for observed outcomes. Intuitively, if we set $\delta = 0$, our estimates would be deeply downward biased because we would effectively say that inequality of opportunity was non-existent. The more weight we put on (2), the more accurate our estimates of type means, i.e. the less the out-of sample-variance in our estimates of the type means. Intuitively, with $\delta = 1$ we would say that we did not care about the precision of our estimates at all, which is the standard practice in today's inequality of opportunity estimations. This instills overfitting and an upward bias in inequality of opportunity estimates. Hence by giving equal weight to both components, the MSE balances upwards and downward biases in inequality of opportunity estimations and thus is a sensible criterion for model evaluation in this context.

A.2 Descriptive Statistics

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

		Sex		Birth area		Presence parents		Household members			Tenancy
Country	Eq. income	Male	Female	Native	EU	Both	One	Adults	Working adults	Children	Owner
AT	25,451	0.499	0.501	0.790	0.070	0.856	0.017	2.73	1.76	2.60	0.585
BE	23,291	0.502	0.498	0.824	0.076	0.855	0.019	2.38	1.59	2.78	0.750
BG	3,714	0.500	0.500	0.994	0.001	0.904	0.012	2.44	2.01	2.07	0.910
CH	42,208	0.495	0.505	0.684	0.197	0.837	0.017	2.55	1.90	2.53	0.546
CY	21,058	0.475	0.525	0.787	0.096	0.900	0.015	2.64	1.67	2.70	0.784
CZ	9,006	0.492	0.508	0.964	0.026	0.851	0.013	2.09	1.92	2.24	0.597
DE	22,221	0.504	0.496	0.868	0.000	0.830	0.020	2.24	1.68	2.32	0.499
DK	32,027	0.495	0.505	0.923	0.026	0.809	0.027	2.22	2.31	2.24	0.736
EE	6,922	0.475	0.525	0.847	0.000	0.756	0.011	2.10	1.80	2.09	0.859
EL	13,184	0.502	0.498	0.890	0.025	0.931	0.019	2.31	1.56	2.33	0.834
ES	17,088	0.505	0.495	0.834	0.051	0.893	0.012	2.88	2.11	2.43	0.819
FI	27,517	0.501	0.499	0.954	0.018	0.829	0.016	2.36	1.75	2.30	0.772
FR	24,299	0.491	0.509	0.885	0.036	0.820	0.022	2.47	1.66	1.75	0.630
HR	6,627	0.499	0.501	0.875	0.017	0.874	0.020	2.56	1.35	2.31	0.902
HU	5,327	0.483	0.517	0.988	0.008	0.844	0.041	2.14	1.75	2.27	0.830
IE	24,867	0.476	0.524	0.783	0.149	0.893	0.078	3.17	3.20	3.20	0.727
IS	22,190	0.493	0.507	0.920	0.042	0.899	0.012	2.42	1.90	2.63	0.893
IT	18,786	0.498	0.502	0.880	0.040	0.901	0.011	2.59	1.62	2.41	0.685
LT	4,774	0.479	0.521	0.939	0.004	0.846	0.016	2.32	2.02	2.46	0.698
LU	37,911	0.501	0.499	0.480	0.401	0.868	0.020	2.53	1.64	2.71	0.734
LV	5,334	0.480	0.520	0.865	0.000	0.763	0.012	1.97	1.76	2.28	0.455
MT	13,006	0.503	0.497	0.944	0.000	0.932	0.020	3.02	1.84	2.68	0.576
NL	25,210	0.491	0.509	0.903	0.020	0.882	0.016	2.10	1.54	3.25	0.575
NO	43,260	0.489	0.511	0.907	0.041	0.913	0.014	2.02	1.76	1.87	0.922
PL	6,103	0.496	0.504	0.999	0.000	0.889	0.015	2.70	1.96	2.44	0.644
PT	10,781	0.494	0.506	0.906	0.022	0.854	0.017	2.68	2.23	2.68	0.544
RO	2,562	0.494	0.506	0.999	0.000	0.919	0.009	2.77	1.90	2.27	0.861
SE	26,346	0.507	0.493	0.846	0.050	0.820	0.035	2.07	1.78	2.35	0.757
SI	13,772	0.504	0.496	0.876	0.000	0.855	0.019	2.53	1.77	2.20	0.746
SK	7,304	0.481	0.519	0.987	0.010	0.920	0.010	2.52	2.08	2.34	0.694
UK	25,936	0.493	0.507	0.848	0.042	0.825	0.024	2.34	2.24	2.41	0.649

Note: Omitted categories are: "Outside Europe" for birth area and "None/Collective house" for the presence of parents, and "Not owned" for the tenancy variable.

Table A.2: Descriptive Statistics (Fathers)

	Birth area		Citizenship		Education			Activity					Main occupation ISCO-08 1-digit								Superv.
Country	Native	EU	Resid.	EU	Prim.	Sec.	Tert.	Empl.	Self-empl.	Unempl.	Retired	House work	1	2	3	4	5 & 0	6	7	8	Yes
AT	0.743	0.093	0.777	0.068	0.007	0.398	0.421	0.714	0.215	0.003	0.010	0.072	0.085	0.063	0.284	0.145	0.147	0.051	0.064	0.046	0.338
BE	0.748	0.100	0.762	0.093	0.016	0.491	0.199	0.699	0.179	0.007	0.011	0.130	0.041	0.127	0.209	0.057	0.054	0.084	0.104	0.126	0.278
BG	0.933	0.004	0.936	0.001	0.029	0.466	0.333	0.899	0.028	0.005	0.004	0.078	0.142	0.207	0.216	0.135	0.058	0.029	0.047	0.065	0.093
CH	0.588	0.286	0.603	0.280	0.051	0.227	0.487	0.653	0.292	0.001	0.003	0.055	0.054	0.077	0.223	0.111	0.065	0.057	0.140	0.131	0.397
CY	0.803	0.082	0.808	0.094	0.045	0.667	0.178	0.566	0.381	0.004	0.009	0.053	0.125	0.122	0.245	0.161	0.109	0.029	0.074	0.071	0.229
CZ	0.878	0.065	0.910	0.036	0.003	0.602	0.216	0.891	0.017	0.001	0.006	0.094	0.053	0.195	0.305	0.039	0.051	0.036	0.125	0.070	0.233
DE	0.800	0.200	0.855	0.145	0.004	0.125	0.496	0.819	0.123	0.008	0.013	0.062	0.040	0.154	0.266	0.059	0.061	0.051	0.158	0.104	0.299
DK	0.935	0.025	0.970	0.020	0.000	0.368	0.418	0.708	0.272	0.004	0.014	0.021	0.009	0.072	0.288	0.160	0.103	0.043	0.070	0.122	0.447
EE	0.603	0.270	0.637	0.233	0.000	0.300	0.338	0.823	0.006	0.003	0.006	0.177	0.053	0.253	0.221	0.034	0.027	0.014	0.053	0.092	0.153
EL	0.887	0.016	0.911	0.015	0.042	0.587	0.135	0.449	0.517	0.002	0.004	0.034	0.055	0.099	0.210	0.308	0.060	0.087	0.026	0.047	0.182
ES	0.836	0.047	0.846	0.046	0.052	0.762	0.064	0.702	0.219	0.006	0.016	0.081	0.137	0.113	0.191	0.145	0.101	0.055	0.076	0.045	0.191
FI	0.827	0.007	0.827	0.007	0.019	0.491	0.182	0.592	0.209	0.016	0.009	0.253	0.044	0.138	0.146	0.135	0.053	0.016	0.085	0.089	
FR	0.789	0.078	0.857	0.057	0.040	0.695	0.073	0.753	0.170	0.003	0.006	0.079	0.223	0.055	0.155	0.103	0.050	0.072	0.111	0.068	0.335
HR	0.822	0.006	0.834	0.004	0.006	0.464	0.312	0.763	0.103	0.037	0.019	0.137	0.228	0.103	0.214	0.049	0.079	0.036	0.088	0.041	0.129
HU	0.962	0.017	0.969	0.012	0.017	0.599	0.241	0.892	0.043	0.001	0.011	0.064	0.137	0.193	0.279	0.094	0.067	0.017	0.052	0.060	0.117
IE	0.792	0.107	0.758	0.094	0.014	0.574	0.258	0.659	0.221	0.049	0.009	0.120	0.158	0.065	0.149	0.155	0.092	0.022	0.042	0.092	0.344
IS	0.918	0.050	0.923	0.044	0.001	0.334	0.486	0.638	0.332	0.001	0.001	0.031	0.042	0.094	0.220	0.180	0.096	0.024	0.076	0.121	0.570
IT	0.823	0.022	0.827	0.020	0.030	0.708	0.136	0.614	0.244	0.016	0.016	0.143	0.118	0.105	0.227	0.099	0.082	0.057	0.074	0.040	0.199
LT	0.899	0.004	0.926	0.004	0.014	0.538	0.228	0.916	0.011	0.000	0.004	0.076	0.214	0.179	0.241	0.080	0.030	0.017	0.038	0.074	0.110
LU	0.387	0.467	0.400	0.466	0.037	0.484	0.316	0.757	0.174	0.001	0.009	0.070	0.039	0.183	0.228	0.112	0.046	0.048	0.118	0.093	0.251
LV	0.572	0.248	0.642	0.165	0.002	0.381	0.297	0.767	0.005	0.002	0.008	0.229	0.083	0.218	0.199	0.069	0.036	0.010	0.037	0.083	0.070
MT	0.952	0.041	0.953	0.040	0.164	0.561	0.180	0.717	0.214	0.013	0.011	0.073	0.106	0.099	0.244	0.050	0.169	0.045	0.106	0.046	0.225
NL	0.829	0.028	0.888	0.022	0.008	0.376	0.285	0.726	0.173	0.006	0.003	0.103	0.031	0.079	0.200	0.086	0.084	0.051	0.155	0.124	0.310
NO	0.897	0.046	0.908	0.041	0.005	0.328	0.390	0.712	0.255	0.002	0.014	0.032	0.032	0.100	0.227	0.111	0.075	0.029	0.167	0.110	0.285
PL	0.955	0.012	0.980	0.003	0.004	0.462	0.448	0.701	0.238	0.002	0.005	0.064	0.078	0.157	0.254	0.237	0.053	0.025	0.053	0.044	0.111
PT	0.932	0.006	0.945	0.006	0.193	0.700	0.031	0.650	0.248	0.002	0.014	0.102	0.077	0.114	0.264	0.185	0.082	0.038	0.060	0.032	0.190
RO	0.938	0.001	0.939	0.001	0.017	0.726	0.088	0.642	0.237	0.004	0.014	0.139	0.104	0.121	0.249	0.253	0.040	0.016	0.034	0.040	0.045
SE	0.945	0.022	0.851	0.061	0.000	0.422	0.350	0.745	0.211	0.002	0.014	0.192	0.019	0.108	0.230	0.086	0.105	0.031	0.067	0.118	0.337
SI	0.769	0.200	0.000	0.000	0.001	0.684	0.166	0.773	0.099	0.013	0.020	0.128	0.173	0.080	0.257	0.089	0.059	0.037	0.100	0.052	0.242
SK	0.935	0.020	0.945	0.011	0.001	0.362	0.497	0.921	0.011	0.002	0.005	0.071	0.128	0.209	0.285	0.030	0.052	0.028	0.095	0.060	0.145
UK	0.800	0.064	0.869	0.039	0.033	0.508	0.228	0.795	0.147	0.025	0.009	0.059	0.083	0.133	0.236	0.036	0.091	0.040	0.085	0.142	0.398

Note: Omitted categories are: "European outside EU" for birth area, "Not Europe" for citizenship, "Illiterate" for education, "Unknown/Dead" and "Other inactive" for activity. ISCO-08 occupation definitions are: 1 "Elementary", 2 "Plant Operator" 3, "Craft/Trades" 4, "Agriculture" 5, "Service and Army" 6, "Clerical", 7 "Technician", 8 "Professional", 9 "Manager"; "Dead/Unknown/Not working" is not shown. Omitted categories for supervisory are: "No" and "Dead/Unknown/Not working".

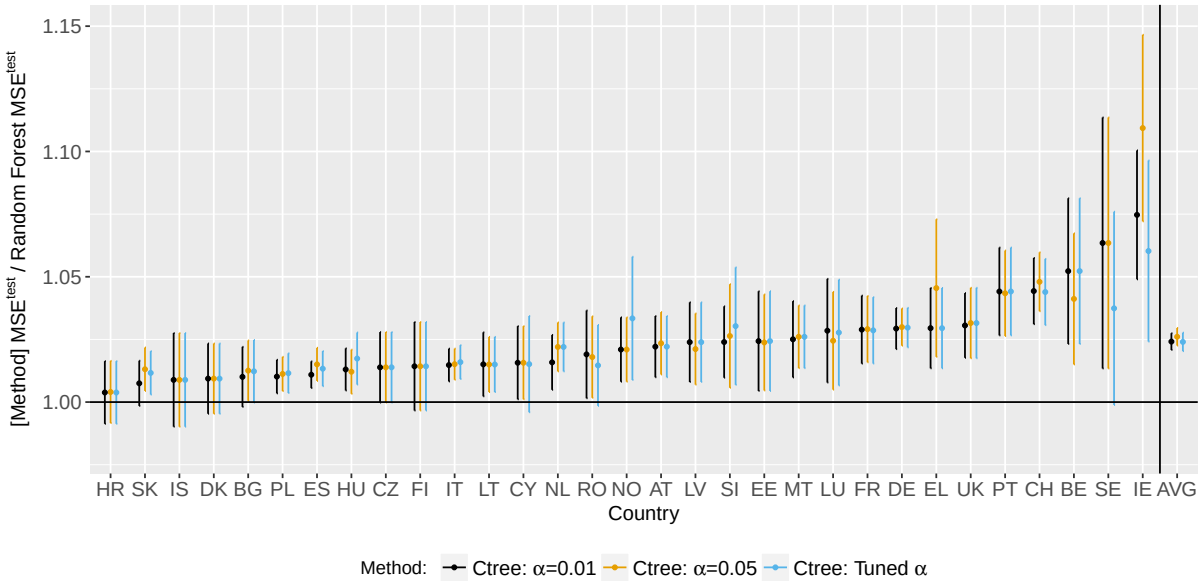
Table A.3: Descriptive Statistics (Mothers)

	Birth area		Citizenship		Education			Activity					Main occupation ISCO-08 1-digit								Superv.
Country	Native	EU	Resid.	EU	Prim.	Sec.	Tert.	Empl.	Self-empl.	Unempl.	Retired	House work	1	2	3	4	5 & 0	6	7	8	Yes
AT	0.740	0.096	0.789	0.065	0.026	0.587	0.328	0.369	0.169	0.002	0.005	0.463	0.087	0.010	0.045	0.128	0.155	0.071	0.009	0.024	0.092
BE	0.755	0.097	0.790	0.092	0.030	0.564	0.201	0.320	0.117	0.006	0.002	0.651	0.069	0.024	0.016	0.002	0.046	0.058	0.045	0.081	0.034
BG	0.931	0.003	0.981	0.002	0.039	0.464	0.357	0.878	0.026	0.007	0.003	0.101	0.152	0.064	0.099	0.181	0.140	0.092	0.040	0.123	0.030
CH	0.567	0.307	0.599	0.286	0.078	0.410	0.399	0.382	0.152	0.001	0.001	0.466	0.068	0.025	0.039	0.055	0.125	0.069	0.069	0.056	0.064
CY	0.804	0.080	0.812	0.091	0.088	0.684	0.162	0.325	0.166	0.001	0.001	0.509	0.220	0.042	0.022	0.036	0.067	0.037	0.020	0.045	0.048
CZ	0.882	0.061	0.946	0.037	0.005	0.670	0.261	0.898	0.007	0.003	0.002	0.096	0.139	0.080	0.105	0.074	0.160	0.149	0.104	0.080	0.088
DE	0.811	0.189	0.862	0.138	0.010	0.284	0.475	0.482	0.050	0.009	0.004	0.493	0.033	0.087	0.015	0.025	0.116	0.089	0.079	0.051	0.059
DK	0.922	0.029	0.935	0.023	0.000	0.531	0.283	0.630	0.069	0.006	0.012	0.321	0.001	0.026	0.052	0.035	0.225	0.123	0.095	0.103	0.122
EE	0.601	0.272	0.726	0.250	0.001	0.334	0.391	0.906	0.004	0.001	0.004	0.092	0.113	0.124	0.051	0.084	0.110	0.097	0.109	0.169	0.085
EL	0.888	0.016	0.916	0.016	0.078	0.592	0.133	0.193	0.277	0.001	0.004	0.532	0.049	0.021	0.034	0.223	0.048	0.039	0.004	0.027	0.026
ES	0.836	0.046	0.849	0.046	0.082	0.802	0.048	0.186	0.069	0.001	0.003	0.748	0.071	0.009	0.021	0.028	0.059	0.021	0.010	0.025	0.029
FI	0.826	0.007	0.933	0.006	0.019	0.559	0.238	0.658	0.204	0.019	0.006	0.151	0.202	0.057	0.046	0.048	0.145	0.122	0.091	0.126	
FR	0.806	0.067	0.880	0.047	0.063	0.724	0.079	0.454	0.085	0.001	0.001	0.463	0.109	0.005	0.049	0.059	0.108	0.111	0.050	0.036	0.072
HR	0.823	0.008	0.848	0.003	0.017	0.634	0.189	0.352	0.053	0.027	0.011	0.596	0.122	0.013	0.034	0.022	0.070	0.046	0.036	0.058	0.033
HU	0.964	0.016	0.980	0.012	0.025	0.655	0.243	0.729	0.022	0.001	0.007	0.252	0.167	0.087	0.075	0.061	0.118	0.113	0.063	0.049	0.044
IE	0.787	0.114	0.761	0.103	0.011	0.546	0.324	0.253	0.048	0.007	0.000	0.700	0.060	0.007	0.014	0.017	0.059	0.052	0.007	0.061	0.082
IS	0.905	0.059	0.924	0.046	0.002	0.626	0.275	0.598	0.102	0.001	0.000	0.305	0.130	0.013	0.028	0.064	0.180	0.109	0.045	0.095	0.149
IT	0.820	0.024	0.862	0.024	0.042	0.779	0.112	0.224	0.080	0.005	0.005	0.698	0.062	0.022	0.031	0.035	0.051	0.029	0.022	0.038	0.041
LT	0.902	0.002	0.959	0.003	0.014	0.519	0.316	0.867	0.014	0.001	0.001	0.124	0.293	0.034	0.112	0.067	0.110	0.049	0.046	0.129	0.068
LU	0.374	0.483	0.393	0.485	0.074	0.587	0.245	0.318	0.106	0.000	0.004	0.579	0.108	0.024	0.015	0.054	0.061	0.036	0.046	0.049	0.047
LV	0.585	0.234	0.793	0.182	0.006	0.414	0.399	0.891	0.003	0.002	0.007	0.106	0.221	0.023	0.093	0.085	0.122	0.098	0.084	0.138	0.074
MT	0.950	0.043	0.957	0.038	0.150	0.652	0.145	0.073	0.015	0.001	0.002	0.919	0.010	0.009	0.004	0.002	0.018	0.009	0.007	0.019	0.011
NL	0.829	0.027	0.907	0.023	0.012	0.532	0.288	0.282	0.056	0.003	0.000	0.665	0.060	0.008	0.011	0.016	0.089	0.052	0.038	0.050	0.037
NO	0.877	0.048	0.891	0.043	0.014	0.368	0.437	0.623	0.106	0.008	0.016	0.270	0.091	0.026	0.017	0.053	0.214	0.114	0.142	0.041	0.065
PL	0.957	0.010	0.990	0.004	0.004	0.524	0.410	0.518	0.261	0.008	0.002	0.226	0.118	0.018	0.080	0.262	0.097	0.071	0.053	0.057	0.050
PT	0.928	0.008	0.950	0.007	0.283	0.631	0.029	0.359	0.197	0.003	0.010	0.444	0.145	0.032	0.059	0.158	0.075	0.025	0.017	0.031	0.048
RO	0.936	0.001	0.939	0.001	0.023	0.728	0.112	0.370	0.219	0.005	0.010	0.440	0.080	0.040	0.076	0.218	0.060	0.026	0.024	0.034	0.010
SE	0.942	0.024	0.855	0.058	0.000	0.409	0.369	0.731	0.058	0.002	0.007	0.582	0.035	0.021	0.009	0.016	0.152	0.057	0.033	0.087	0.095
SI	0.791	0.178	0.000	0.000	0.004	0.752	0.148	0.578	0.071	0.005	0.010	0.351	0.193	0.006	0.066	0.061	0.091	0.085	0.093	0.047	0.089
SK	0.932	0.023	0.980	0.010	0.001	0.451	0.482	0.846	0.006	0.004	0.002	0.153	0.203	0.052	0.096	0.034	0.161	0.107	0.110	0.075	0.048
UK	0.808	0.064	0.877	0.036	0.042	0.679	0.099	0.577	0.051	0.087	0.003	0.375	0.127	0.044	0.028	0.005	0.152	0.078	0.068	0.097	0.104

Note: Omitted categories are: "European outside EU" for birth area, "Not Europe" for citizenship, "Illiterate" for education, "Unknown/Dead" and "Other inactive" for activity. ISCO-08 occupation definitions are: 1 "Elementary", 2 "Plant Operator" 3, "Craft/Trades" 4, "Agriculture" 5, "Service and Army" 6, "Clerical", 7 "Technician", 8 "Professional", 9 "Manager"; "Dead/Unknown/Not working" is not shown. Omitted categories for supervisory are: "No" and "Dead/Unknown/Not working".

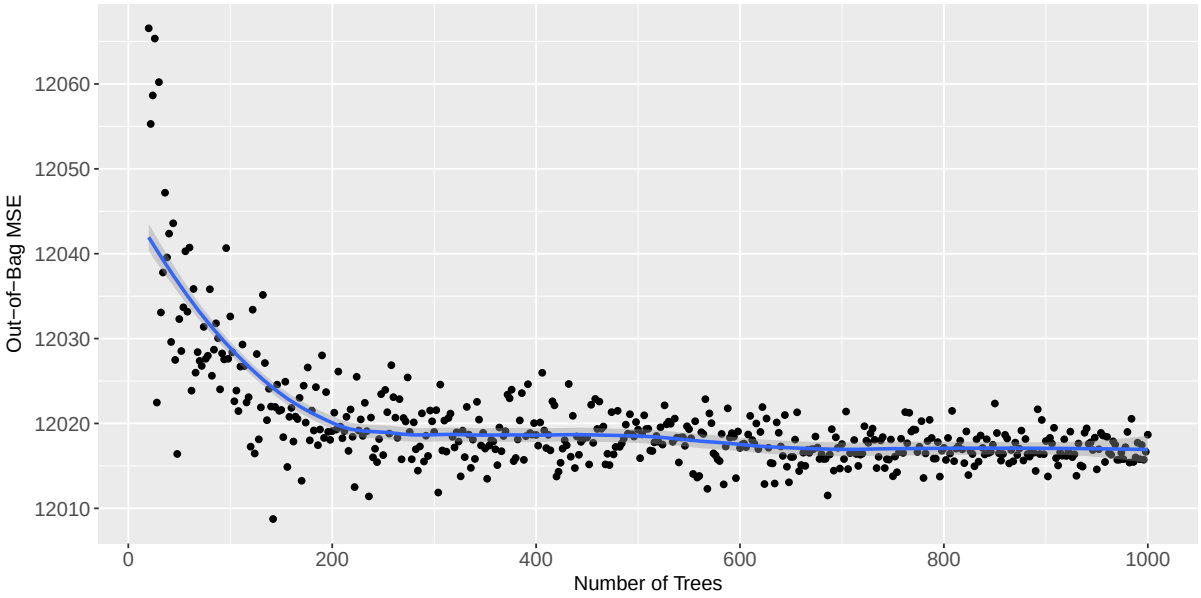
A.3 Empicial Robustness Checks

Figure A.1: Tuning Conditional Inference Trees



Note: The figure compares the test error of random forests with different conditional inference trees. The test error of random forests is standardized to equal 1, such that a test error larger than 1 represent worse fits than random forests. "Ctree: tuned α" uses cross-validation to tune α. 95% confidence intervals are derived based on 200 bootstrapped re-samples of the test data using the normal approximation method.

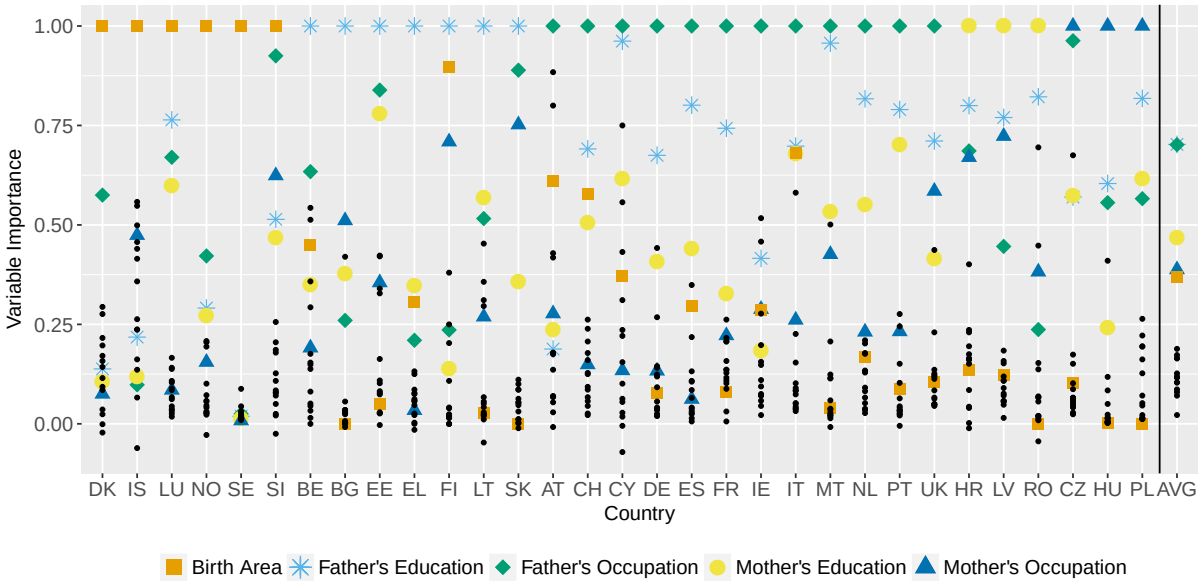
Figure A.2: Optimal Size of Forests



Note: The figure compares the MSE^{OOB} for Germany using varying forest sizes (different levels of B). We allow for 6 circumstances to be considered at each splitting point ($\bar{P} = 6$). The blue line indicates the loss of fit. After around 200 trees, improvements in the error tend to be negligible. Similar patterns were found with other countries and other levels of $\bar{P}$. For this reason, we set $B^* = 200$ in our random forests.

A.4 Opportunity Structures

Figure A.3: Variable Importance Plot



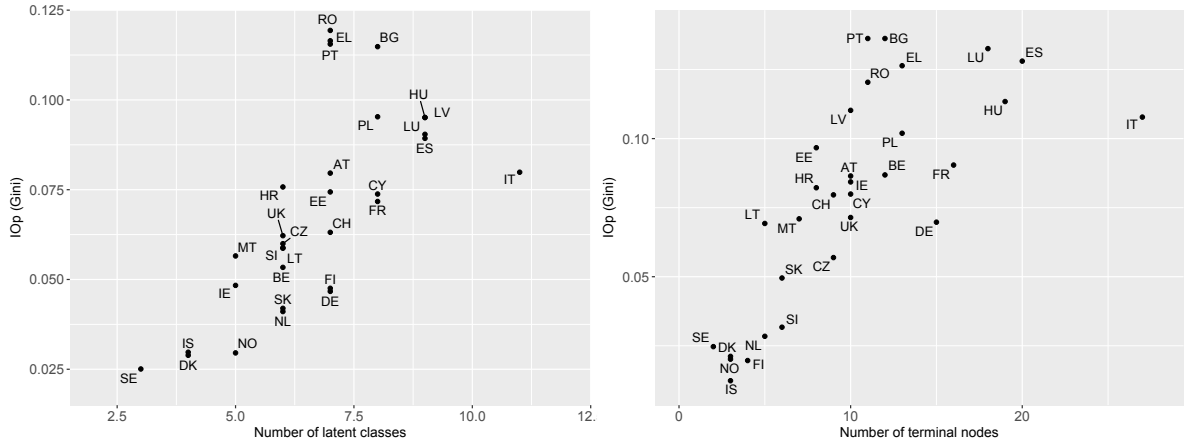
Note: Each dot shows the importance of a particular circumstance for the predictions from our random forest. The importance of a circumstance is measured by permuting the circumstance, calculating a new MSE^{OOB}, and computing the difference in the MSE^{OOB} between the original model and the model with the permuted circumstance. The importance measure is standardized such that the circumstance with the greatest importance in each country equals one.

A.5 Tree Complexity

Figure A.4: Correlation of Complexity and Magnitude of Estimate

(a) Latent Class Analysis

(b) Conditional Inference Tree



Note: Complexity of the opportunity structure is proxied by the number of latent classes and the number of terminal nodes, respectively.