Brunori, Paolo; Neidhöfer, Guido

Working Paper

The evolution of inequality of opportunity in Germany: A machine learning approach

Documento de Trabajo, No. 259

Provided in Cooperation with:
Centro de Estudios Distributivos, Laborales y Sociales (CEDLAS), Universidad Nacional de La Plata

Suggested Citation: Brunori, Paolo; Neidhöfer, Guido (2020) : The evolution of inequality of opportunity in Germany: A machine learning approach, Documento de Trabajo, No. 259, Universidad Nacional de La Plata, Centro de Estudios Distributivos, Laborales y Sociales (CEDLAS), La Plata

This Version is available at:
https://hdl.handle.net/10419/250348

CEDLAS
Facultad de Ciencias Económicas
UNIVERSIDAD NACIONAL DE LA PLATA

DOCUMENTOS DE TRABAJO

The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach

Paolo Brunori y Guido Neidhofer

Documento de Trabajo Nro. 259
Marzo, 2020
ISSN 1853-0168
www.cedlas.econo.unlp.edu.ar

Cita sugerida: Brunori P. y G. Neidhofer (2020). The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach. Documentos de Trabajo del CEDLAS Nº 259, Marzo, 2020, CEDLAS-Universidad Nacional de La Plata.

The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach*

Paolo Brunori†, Guido Neidhöfer‡

February 12, 2020

Abstract

We show that measures of inequality of opportunity (IOP) fully consistent with Roemer (1998)'s IOP theory can be straightforwardly estimated by adopting a machine learning approach, and apply our novel method to analyse the development of IOP in Germany during the last three decades. Hereby, we take advantage of information contained in 25 waves of the Socio-Economic Panel. Our analysis shows that in Germany IOP declined immediately after reunification, increased in the first decade of the century, and slightly declined again after 2010. Over the entire period, at the top of the distribution we always find individuals that resided in West-Germany before the fall of the Berlin Wall, whose fathers had a high occupational position, and whose mothers had a high educational degree. East-German residents in 1989, with low educated parents, persistently qualify at the bottom.

Keywords: Inequality, Opportunity, SOEP, Germany.
JEL Classification: D63, D30, D31

*We are grateful to Torsten Hothorn, Paul Hufe, Daniel G. Mahler, Julia Heigle, Eduard Brüll, Robin Jessen, Friedhelm Pfeiffer, and John Roemer for useful suggestions. Furthermore, we would like to thank participants of presentations given in Berlin, Mannheim, Milan, Paris, and Trento for their comments. Guido Neidhöfer gratefully acknowledges funding from the state Baden-Württemberg within the SEEK program. Any errors remain our own.

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

‡ZEW - Leibniz Centre for European Economic Research, L7 1 - 68161 Mannheim, Germany, guido.neidhoefer@zew.de.

1 Introduction

The ideal of equality of opportunity has fascinated mankind for centuries. Its popularity among people from both sides of the political spectrum probably derives from the fact that it encompasses and balances two aspects; equality of outcomes and freedom of choice. In addition, while the normative evaluation of outcome inequality is controversial, nobody would argue against equality of opportunity as an important goal. At the same time, the political rhetoric demanding it might be sufficiently vague that it allows for different interpretations.

For a long time, moral philosophers and welfare economists defined and conceptualized the notion of equality of opportunity as well as its implications from a normative point of view. Rawls (1971) proposed a theory of social justice in which redistribution of outcomes and social roles was somehow limited by the need to take into consideration individuals' responsibility. Dworkin (1981) went a step further, focusing on the distinction between preferences and resources. From his perspective, inequality in final conditions is morally objectionable, and calls for redistribution in the case that these differences arise from unequal resources, though not when they arise from preferences and choices. Among economists the most influential formalization of the principle of equal opportunity is due to Roemer (1998). Roemer's definition of inequality of opportunity comprises the interplay between circumstances that individuals are exposed to and the degree of effort they exert; circumstances are exogenous factors outside individual control, while effort indicates the result of choices for which the society wishes to hold individuals responsible.

Roemer's theory of equal opportunity has triggered a lively empirical literature. Metrics were proposed to measure inequality of opportunity based on the distance of a given distribution of individual outcomes from equal opportunity (among others, Almås et al., 2011; Checchi and Peragine, 2010; Lefranc et al., 2009). A popular approach is, for instance, the regression-based method proposed by Ferreira and Gignoux (2011) to quantify the share of total inequality due to opportunity. Two recurrent issues of the empirical literature are hereby: *i.* the need to identify the set of circumstances beyond individual control and *ii.* the need to assume how these circumstances interact with effort in determining individual outcomes. Both choices have been shown to crucially affect the estimated level of inequality of opportunity (Brunori et al., 2018b).

Recent contributions have proposed approaches to improve the empirical specification of the underlying models, finding consistent econometric methods to identify the relevant circumstances, and eventually estimate inequality of opportunity. Within this literature, Li Donni et al. (2015) and Brunori et al. (2018a) propose data-driven approaches to identify Roemerian types (i.e. sets of individuals characterized by identical circumstances). Still, besides the identification of circumstances, in the vast majority of empirical contributions so far inequality of opportunity is estimated without taking into consideration the role of effort, which is key in Roemer's theory (Ramos and va de Gaer, 2019) .

In this study, we propose a method that builds on a data-driven approach and exploits two machine learning algorithms (namely, regression trees and polynomial approximation) to estimate inequality of opportunity consistent with Roemer's original theory. In a first step, we follow Brunori et al. (2018a) to identify types and estimate opportunity trees. Then, we develop a new approach to estimate the degree of effort by polynomial approximation of the conditional distribution of household income for each type. This enables us to precisely estimate the relationship between effort and outcome, even for types with a small sample size.

We apply this novel approach to estimate the evolution of inequality of opportunity in Germany from shortly after the fall of the Berlin wall to the present. Germany is an interesting case study for our analysis because of the societal changes that the country underwent during the last thirty years. Our application using the Socio-Economic Panel (SOEP) shows that the opportunity structure of the German society is much more complex today than it was back in the nineties. The number of types identifiable in SOEP has increased substantially over time;

a pattern that is not fully explained just by changes in survey characteristics. Besides the increasing societal diversity, having being located in the Eastern or Western part of Germany in 1989 is, more than two decades after reunification, constantly a significant circumstance defining the subdivision in types over the course of time.

Our analysis uncovers another interesting peculiarity. While usually societies characterized by a large number of types tend to have much higher levels of inequality of opportunity, this is not the case for Germany over time. Controlling for the sample size, albeit the number of types increases by roughly 75%, the level of inequality of opportunity in 2016 is just around 7% higher than in 1992. Generally, during this period Germany experienced first a slow decrease in inequality of opportunity after reunification, and then a sudden rise coinciding with rising income inequality and the implementation of the *Hartz-reforms*, a set of substantial changes to the German labour market and welfare benefits system that had persistent repercussions for German society. Thenceforth, inequality of opportunity stayed at this relatively high level and followed a stable trend, with a slight decrease from 2010 onward.

2 Inequality of Opportunity

Roemer (1998) represents the seminal contribution for the empirical literature on inequality of opportunity (IOP). In his book, Roemer did not explicitly write down a definition of inequality of opportunity. Rather, his theory proposes a criterion to select the redistributive policy that would equalize opportunity in a society. This theory has been translated into more than one definition of inequality of opportunity.

Roemer's theory distinguishes between two categories of factors that determine individual outcomes: factors over which individuals have control, which he calls *effort*, and factors for which individuals cannot be held responsible, which he calls *circumstances*. He defines equal opportunity in the distribution of a certain desirable outcome as the scenario in which individuals are compensated for the difference in their circumstances, insofar as those differences affect the advantage they attain. To realize equal opportunity, Roemer proposes a partition of the population into *types*. A type is a set of individuals characterised by exactly the same circumstances (gender, race, socioeconomic background,...). When exerting effort, individuals in the same type have the same ability to transform resources into outcomes. Therefore, an equal opportunity policy prescribes to ignore within-type variability in outcomes, which by definition is due to individual effort, and requires the removal of any between-type inequality .

Roemer's definition of equal opportunity can be formalized in a simple model. In a population of $1, ..., N$ individuals, individual i obtains an outcome of interest, y_i, as the result of two sets of traits: a set of circumstances beyond her control, $\mathbf{C}_i$,and a responsibility variable, e_i, called effort.

$$y_i = g(\mathbf{C}_i, e_i), \ \ \forall i = 1, ...N$$

$\mathbf{C}_i$ contains $J > 0$ circumstances, each circumstance, $C^j \in \mathbf{C}$, is characterized by a total of x^j possible realizations. All possible combinations of realizations taken one at a time from $\mathbf{C}$ define a partition of the population into types. This partition is made of a maximum of $K = \Pi_{j=1}^{J} x^j$ non-empty subsets, where every individual is included in one and only one of the subsets. Note that this simple model does not introduce any random component or uncertainty.

Equality of opportunity is realized when individuals exerting same effort obtain the same valuable outcome, independently from the type they belong to. In order to measure to what

extent this principle is violated, one has to compare the outcome of individuals belonging to different types but exerting the same effort. Because effort is typically unobservable, Roemer proposes a method for identifying effort. His method is based on three assumptions. First, we fully observe relevant circumstances, that is we correctly assign individuals to types. Second, the outcome is assumed to be monotonically increasing in effort: in every types higher effort implies higher outcome.

$$y^k(e_i) \geq y^k(e_j) \iff e_i^k \geq e_j^k, \ \forall k = 1, ..., K, \forall e_i, e_j \in \mathbb{R} \tag{1}$$

Third, the degree of effort exerted is by definition a variable orthogonal to circumstances. In Roemer's view, if individuals belonging to different types face different incentives and constraints in exerting effort, this is to be considered a characteristic of the type and therefore included among circumstances beyond individual control.

For example, a student with well educated parents may find it much easier to spend hours sitting at her desk, while a student growing up in a less favourable environment may find it harder to study. Roemer believes that the distribution of effort is, indeed, a characteristic of the type: *thus, in comparing efforts of individuals in different types, we should somehow adjust for the fact that those efforts are drawn from distributions which are different, a difference for which individuals should not be held responsible.* Roemer (2002) p. 458.

Roemer therefore distinguishes between the 'level of effort' and the 'degree of effort' exerted by an individual. The latter is the morally relevant variable of effort, and is identified with the quantile of the effort distribution for the type to which the individual belongs. We denote with $G^k(e)$ the distribution of effort within type k and with $\pi \in [0, 1]$ its quantiles.

If effort is not observable, but the outcome is monotonically increasing in e, Roemer suggests to identify the degree of effort exerted by a given individual with the quantile of the type-specific outcome distribution she sits at: $y^k(G^k(e)) = y^k(\pi)$. This definition of effort is insensitive to differences in the absolute level of effort exerted that, in Roemer's view, are due to circumstances beyond individual control, and it permits the comparison of effort exerted by individuals in different types.

Then the requirement of same outcome for individuals exerting same effort can be rewritten in terms of type-specific outcome distributions:

$$y^k(\pi) = y^l(\pi) \iff F^k(y) = F^l(y), \forall \pi \in [0, 1]; \quad k, l = 1, ..., K \tag{2}$$

Where $F^k(y)$ is the type-specific cumulative distribution of outcome in type k.

A measure of inequality of opportunity quantifies to what extent this equality of opportunity principle is violated. This is done measuring the variability of the outcome distribution within individuals exerting same effort (Almås et al., 2011; Checchi and Peragine, 2010; Ferreira and Gignoux, 2011; Lefranc et al., 2009). By construction, these measures take value zero when equation (2) is satisfied and all individuals exerting the same effort obtain the same outcome, and increase with larger differences in outcomes obtained by individuals exerting the same degree of effort.

For example, the ex-post measure of inequality of opportunity proposed by Checchi and Peragine (2010) evaluates inequality in the standardized distribution $\tilde{Y}_{EP}$ obtained replacing individual outcome with:

$$\tilde{y}_i^k(\pi) = y_i^k(\pi)\frac{\mu}{\mu^\pi}, \quad \forall i = 1, ..., N; \quad k = 1, ..., K; \quad \forall \pi \in [0, 1]. \tag{3}$$

Where $y_i^k(\pi)$ is the outcome of individual i belonging to type k and sitting at quantile π of the type specific effort distribution, and μ^π is the average outcome of individuals sitting at quantile π across all types, and μ is the population mean outcome. Note that in the standardized

distribution the average value for individuals sitting at all quantiles is the same, that is, between-quantile inequality has been removed. On the contrary, the within-quantile relative distance of outcome is preserved. Inequality of opportunity, IOP_{EP}, is then inequality in the standardized distribution:

$$IOP_{EP} = I(\tilde{Y}_{EP}) \tag{4}$$

Where I is any inequality measure satisfying the typical properties, including scale invariance.[1]

Ex-post measures of inequality of opportunity are not frequently implemented in empirical analysis. The majority of applied studies focus on a second, less demanding, definition of equal opportunity. The *ex-ante* equality of opportunity is a 'weak equality of opportunity' criterion that allows some inequality within groups of individuals exerting the same effort but requires that mean advantage levels should be the same across types (Ferreira and Gignoux, 2011).

The ex-ante measure of inequality of opportunity first proposed by Van de gaer (1993) is a measure based on this weaker definition. The approach interprets the type-specific outcome distribution as the opportunity set of individuals belonging to each type. The (utilitarian) value of the opportunity set of each type is the mean outcome of the type. Therefore, inequality of opportunity in this case is simply between-type inequality, the counterfactual distribution $\tilde{Y}_{EA}$ is obtained replacing individual outcome with:

$$\tilde{y}_i^k(\pi) = \mu^k, \;\; \forall i = 1, ..., N; \quad \forall k = 1, ..., K; \quad \forall \pi \in [0, 1] \tag{5}$$

Where μ^k is the mean outcome of type k.

$$IOP_{EA} = I(\tilde{Y}_{EA}) \tag{6}$$

Adopting the ex-ante approach simplifies the measurement of inequality of opportunity, which becomes equivalent to a measure of between-group inequality. Furthermore, IOP_{EA} is by far the most popular measure of inequality of opportunity.[2] However, this approach implies a loss of consistency with the principle of compensation, which, in its original formulation, is the fundamental ethical principle of Roemer's theory of equal opportunity, stating that individuals exerting same effort should obtain same outcome (see Fleurbaey and Peragine (2013) for a discussion of this incompatibility).[3]

3 Machine Learning Estimation of IOP

The estimation of IOP_{EP} is based on two fundamental tasks: the identification of Romerian types and the measurement of the degree of effort exerted. We adopt a machine learning approach to accomplish both. The partition into types is obtained estimating regression trees; the degree of effort is measured estimating the type specific outcome distribution by a polynomial approximation.

[1] Researchers interested in measuring inequality of opportunity with a translation invariant inequality measure should replace equation (4) with: $\tilde{y}_i^k(\pi) = y_i^k(\pi) + \mu - \mu^j$.

[2] The project Equalchances.org estimated comparable estimates of IOP_{EA} for 51 countries.

[3] Quoting Ramos and va de Gaer (2019): *"Most of the empirical literature continues to treat [ex-ante and ex-post approaches] as interchangeable, by motivating their concern with inequality of opportunity from ex-post intuitions and using ex-ante measures of inequality of opportunity."*

3.1 Identification of types

The first step to estimate both equation (4) and equation (6) is the identification of circumstances beyond individual control that define types.[4] The selection of circumstances is a key aspect of any empirical analysis of IOP because estimates have been shown to be sensitive to the number of types considered (Brunori et al., 2018b; Ferreira and Gignoux, 2011; Rodríguez, 2008).

In principle, one should include all variables beyond individual control that can affect the outcome. However, this is not a realistic option. First, surveys typically contain only a subset of all the exogenous determinants of individual outcomes. Second, even when a rich dataset is available, the sample size constrains the number of circumstances that can be considered if one has to reliably estimate the counterfactual distributions.

The so-called non-parametric approach proposed by Checchi and Peragine (2010) attempts to exactly implement Roemer's definition of types: the partition in types is obtained interacting all circumstances. This typically results in a partition with a large number of types. However, many of these types are sparsely populated making it impossible to estimate with accuracy the type-specific cumulative distribution of the outcome.

Empirical exercises generally address this issue by limiting the number of circumstances used to define types. Also, the categories that describe circumstances are recoded reducing their variability. For instance, districts of birth are aggregated into macro-region; parental occupations become a binary variable for white or blue collar workers; ethnicity becomes a dummy for minorities. The resulting types have large sample sizes but are based on arbitrary choices. These *ad-hoc* methods to identify types severely undermine the interpretability and comparability of estimates of inequality of opportunity (Brunori et al., 2018b).

Two papers have proposed data-driven criteria to identify Romerian types. Li Donni et al. (2015) specify that types are inherently unobservable and suggest grouping individuals in types using latent class models. Latent class models assign individuals to types based on observed circumstances, interpreted as observable manifestations of the underlying latent types. Type membership is determined in the attempt to maximize local independence. Local independence means that, conditional on class membership, observed items are conditionally independent from each other. Once latent types are identified, the researcher can move to the second step of the analysis which consists of identification of the effort exerted.

Latent types are an appealing theoretical construct. However, their implementation has two problematic aspects. First, in latent class models the number of classes is exogenously given. Li Donni et al. (2015) suggest selecting the number of latent types guided by the Bayesian information criterion (BIC). BIC evaluates the likelihood of the model introducing a penalty term for the number of parameters estimated. The BIC selects the most appropriate model balancing between choosing a model able to closely fit the data in the sample, and choosing a model with the lowest possible number of parameters. When estimating models based on an increasing number of latent types BIC will first rise, indicating that additional classes substantially improve model fit, and then, when the effect of the penalty dominates, BIC will start to decline. According to Li Donni et al. (2015) the most appropriate partition is obtained by choosing the number of classes that produces the highest BIC.

A perfect fit is obtained when the distribution of manifest variables is orthogonal to classes, that is, when local independence is fully satisfied. The BIC of a latent class model therefore evaluates the capacity of the model to explain the correlation of manifest variables in the sample. However, when estimating inequality of opportunity the aim is not to explain covariance of circumstances, but to identify the partition in types that best explains the outcome variability.

[4]Note that other methods, based on a parametric estimation of the function $g()$, have been proposed (Bourguignon et al., 2007; Ferreira and Gignoux, 2011). In what follows we limit the discussion to methods that do not impose a functional form on the data generating process and explicitly identify types.

A criterion such as the BIC may not be the most appropriate for selecting the number of latent types. This is a specific case of the more general problem of using latent class membership as predictor for a distal dependent variable. As discussed by Lanza et al. (2013), such an approach is likely to produce attenuated estimates of the effect of the latent class membership on the outcome.

Another issue concerns the choice of observable circumstances considered in the latent class model. As discussed above, the problem of arbitrary selection of circumstances severely undermines the non-parametric estimation of inequality of opportunity; this problem is attenuated but not completely solved when using latent class models. The number of parameters one needs to estimate when applying a latent class model is, in fact, a function of the number of classes, the number of circumstances considered, and the number of values each circumstance can take. This implies that the choice of circumstances considered, which is arbitrary, will affect the result.

This weakness of the latent types approach highlights that a proper method that aims to estimate inequality of opportunity needs to comprise both, the identification of types based on observed circumstances, and a variable selection criterion that would select the most appropriate set of the (possibly) many observable circumstances.

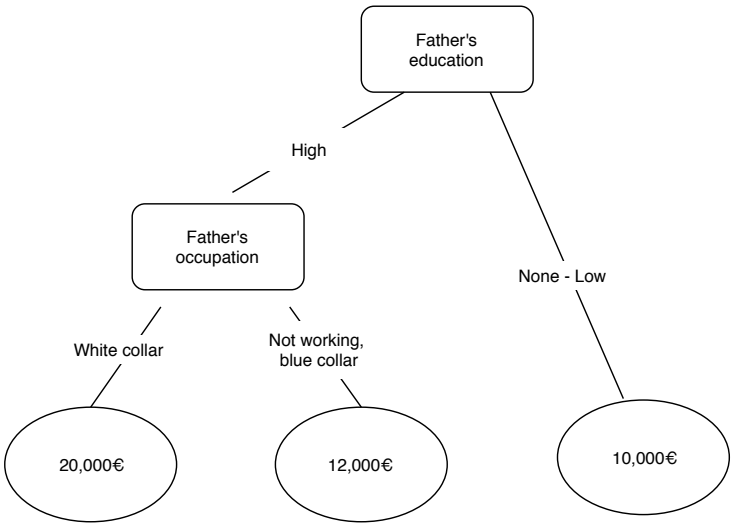
In light of this Brunori et al. (2018a) propose the use of a machine learning algorithm, known as *conditional inference regression trees*, to identify Romerian types. Regression trees are prediction algorithms introduced by Morgan and Sonquist (1963) and popularized by Breiman et al. (1984) almost 20 years later. The algorithm aims to predict an outcome out of sample based on a number of covariates. This is done by partitioning the space of the regressors in non-overlapping regions. The name *tree* comes from the way this algorithm can be graphically represented as an upside-down tree. Figure 1 shows an example of a regression tree for predicting income based on two regressors: parental education and parental occupation. The predicted income is simply the average outcome of individuals assigned to each terminal node (ovals at the bottom of the tree). Regression trees are generally grown in the attempt to maximize the ability of the model to predict out-of-sample. That is, trees aim at maximizing the variability of the dependent variable that can be explained by between-node variability, without overfitting the model. A too deep tree (that is, an overfitted tree) would result in a very low in-sample error but would poorly perform out-of-sample.

Various methods exist for growing trees while avoiding overfitting. *Weakest link pruning* estimates the mean squared error out-of-sample for all trees obtained replacing a sub-tree with a terminal node. This method of pruning is computationally costly in cases with a large set of regressors. Other methods such as *cost complexity pruning* or *conditional inference trees* can be used to prevent trees from growing too deep (James et al., 2013). Conditional inference trees introduced by Hothorn et al. (2006) prevent overfitting by growing the tree while conditioning the splitting on a sequence of statistical tests. The algorithm follows a stepwise procedure:

1. set a confidence level $(1 - \alpha)$, typically 0.99 or 0.95;
2. test all null hypothesis of independence between the individual outcome and each of all the observable circumstances;
3. if none of the hypotheses are rejected with the selected level of confidence, stop;
4. if for one or more circumstances the adjusted p-value[5] of the independence test is below α, select the circumstance with the smallest p-value;
5. for every possible way the selected circumstance can be used to divide the sample into two subgroups, test the hypothesis of same mean outcome in the two resulting subgroups;

[5]The algorithm uses the Bonferroni correction for multiple hypothesis testing.

Figure 1: A regression tree



A simplified example of a regression tree explaining individual income variability. The tree is made of two splitting points (father's education and father's occupation) and three terminal nodes (ovals reporting average outcome for each type).

6. choose as splitting point for partitioning the sample the value (or the two groups of categories) for which the test produces the smallest p-value;

7. repeat the algorithm for each of the resulting subsamples.

The use of this algorithm presents a number of advantages: first, the choice of circumstances used to construct types is no longer arbitrary. Even when very large sets of observable circumstances are available, the algorithm will use only the characteristics that have the strongest association with the outcome. Second, the model specification is no longer exogenously given: how circumstances interact in determining the outcome is driven by the attempt of the algorithm to explain the variability of the outcome. Third, the algorithm automatically provides a test for the null hypothesis of equality of opportunity. Indeed, it is not impossible that the algorithm stops at step 3; the original sample is not split and the tree is made of a single terminal node. In this particular case, we could not reject the null hypothesis of equal opportunity[6]. Fourth, but no less important, opportunity trees tell us a story about the structure of opportunity that is immediately possible to understand even without formal statistical training.

3.2 Identification of effort

Once types are identified, the second step consists of estimating effort. Adopting Roemer's identification strategy this is done by estimating the shape of the type-specific outcome distribution in all types. Previous contributions select an arbitrary number of quantiles, generally not larger than 10, and estimate equation (4) setting μ^{π} equal to the average outcome across all individuals belonging to the $j-$th quantile of their type specific outcome distribution.

Although not explicitly discussed by these contributions, the need to estimate the distribution of outcomes for each type has a clear impact on the empirical exercise. If in the ex-ante approach the main constraint when identifying types is the need to reliably estimate their mean, following the ex-post approach the need to estimate the type-specific outcome distribution for each type

[6]From this point of view, the construction of a conditional inference tree can be interpreted as a robust version of a statistical test for the null hypothesis of equal opportunity in the spirit of Lefranc et al. (2009).

imposes a more severe trade-off. On the one hand, a precise description of the data generating process requires the consideration of a sufficiently large number of types; on the other, the estimation of the type-specific outcome distribution requires a large number of observations in each type. Much larger than the sample size required for estimating a single parameter for each type. Particularly because Roemer's strategy imposes no restriction on the type-specific outcome distribution, the researcher has to estimate a number of parameters equal to the number of quantiles.

As Luongo (2011) clarifies, inequality of opportunity estimates are sensitive to the selected number of quantiles. However, if one can imagine that –theoretically– a precise number of Roemerian types does exist, it is clear that quantiles are used to approximate an intrinsically unknown continuous distribution function. Hence, there is no true number of quantiles.

This paper proposes a non-arbitrary criterion to approximate the type-specific outcome distribution based on a procedure suggested by Hothorn (2018). Such a criterion makes measures of inequality of opportunity *à la* Roemer less dependent on discretionary methodological choices and more easily comparable across time and space.

Moreover, our method explicitly addresses the problem of balancing the need to precisely estimate the distribution of outcome in each type, and the data constraints; constraints that will typically differ across types of different sample size.

We approximate the shape of the type-specific outcome distribution $F^k(y)$ using the Bernstein polynomial, that is a linear combination of Bernstein basis polynomials.[7] The Bernstein basis polynomial of degree m for some positive continuous variable $t \in [a, b]$ is defined as the set of polynomials:

$$\left\{ b_{j,m}(t,a,b) = \frac{1}{(b-a)^m} \begin{pmatrix} m \\ j \end{pmatrix} (t-a)^j (b-t)^{m-j}, \ \ \forall j = 1, ..., m \right\} \tag{7}$$

A linear combination of Bernstein basis polynomial has the form:

$$B_m(t,a,b) = \sum_{i=0}^{m} \beta_i b_{j,m}(t,a,b) \tag{8}$$

In each type we select the degree of the Bernstein polynomial that maximizes the out-of-sample Log likelihood in approximating the real cumulative distribution function. Out-of-sample Log likelihood is estimated by ten-fold cross-validation. The algorithm proceeds for each type $k = 1, ..., K$ with the following steps:

1. partition the population of type k into 10 non-overlapping sets of approximately equal size (folds);
2. for every $b = 1, ..., 10$;
    a. for every fold $f = 1, ..., 10$;
        I. obtain the *training* sample by excluding from the sample the f-th fold that will be later used as *test* sample;
        II. estimate the shape of the type-specific outcome distribution with a monotone increasing Bernstein polynomial of order b on the training set;
        III. predict the cumulative distribution of the type based on the Bernstein coefficients $\hat{F}_b^k(y)$ on the test set;

[7] Introduced in 1912 by Sergei Bernstein, Bernstein polynomials become known as the the mathematical basis of the Bèzier curves, used first to design automobile bodies and, more recently widely adopted in computer graphics to model smooth curves (Farouki, 2012). We opt for this approximation method as it has been shown to outperform competitors, such as kernel estimators, in approximating distribution functions (Leblanc, 2012).

IV. estimate the out-of-sample log-likelihood (LL_b^f);

V. store LL_b^f;

b. we calculate and store $LL_b = \sum_{f=1}^{10} LL_b^f$;

3. select the maximum $LL_{b^*} \in [LL_1, ..., LL_{10}]$;

4. b^* indicates most appropriate order for the Bernstein polynomial approximating the types-specific distribution function of type k.

The algorithm, repeated for all types, produces parametric approximations of the type-specific outcome distribution functions based on the coefficients of Bernstein polynomials of different order. Under Roemer's assumption, the quantiles of such distributions measure the degrees of effort individuals exerted. The estimation of equation (4) can then be conceptualized as inequality between a set of weighted, type-specific outcome distributions, and will depend on population weights and the Bernstein coefficients. Equation (4) can then be precisely approximated using a sufficiently high number of points to approximate each type-specific outcome distribution.

Equation (3) is then estimated by multiplying the outcome of each individual $i = 1, ..., N$ belonging to each type $k = 1, ..., K$ by the average outcome in the population (μ) divided by the expected outcome of an individual exerting the same degree of effort (μ^π), independently from the type she belongs to.

4 Data

4.1 The Socio-Economic Panel

Our analysis on the evolution of inequality of opportunity in Germany, applying the methods explained above, is based on the Socio-Economic Panel (SOEP; see Goebel et al., 2019). The SOEP is a representative longitudinal survey of private households in Germany conducted annually since 1984, including the East German population since 1991. It is also the largest regular survey of immigrants and their offspring in Germany.

The SOEP is one of the main data sources for distributional analysis in Germany. Furthermore, it includes a remarkable amount of retrospective information on individual characteristics that shall be indicative for the circumstances faced in childhood. For instance, questions on the education and occupation of parents, region of birth, migration background, and country of origin are included in the questionnaire and made comparable across the survey years.

We use the v33 version of SOEP including all subsamples apart from the two newly added refugee samples. We restrict the sample to individuals in the age range 30 to 60 with available information on household income and all circumstances that we include in our analysis. Note, the reported income in a survey year refers to the year before the survey was conducted.

Descriptive statistics for the 25 waves used are reported in Appendix A. In the same Appendix, we show for each survey wave the share of individuals with missing information for potential circumstances that we identify based on the past literature on IOP and the specific German context. In order to ensure the maximum possible comparability of estimates across time, we choose the set of circumstances with the lowest number of missing values across all survey waves. The circumstances that we include are: sex, migration background, resident in East or West Germany in 1989, Father's and Mother's education and training, Father's occupation measured by the ISCO-code (one digit), the number of siblings, and an indicator of whether the individual is disabled.

We set 1992 as the first survey wave because information on household income for people in East Germany is available from that year on, hence the analysis starts from the year 1991. To

warrant the representability of our analysis at the national level in every year, each observation is weighted by the inverse probability of selection. Household income is displayed in Euro at 2011 prices. Our outcome of interest is household disposable equivalent income (applying the square root scale). To avoid life-cycle bias in our estimates we compute, for every year, the deviation of individual income from its expected value given the respondent's age. Our outcome of interest is therefore the deviation from what is predicted by a regression in which disposable household equivalent income declared by the respondent is regressed on her age and her age squared. Hence, individuals with outcome higher than one have a larger equivalized income than the average resident in Germany in their age specific reference group.[8] Total inequality trends for this residual outcome measure, as well as total equivalized household income, can be found in Appendix A.

4.2 Sample size

A key challenge for our analysis is that our sample size varies considerably across waves of the SOEP, ranging between 2,868 and 13,160 (see Appendix A). Conditional inference regression trees have been shown to be sensitive to the sample size because the splitting points are conditioned on a sequence of statistical tests and, when sample size is small, p-values of the tests tend to be higher. Other things held constant, the larger the sample size the deeper might be the resulting tree. Deeper trees, made of many terminal nodes, usually tend to produce larger IOP estimates.

In order to overcome this problem and maximize the comparability of estimates over time, we proceed in two steps. First, we show the partition in types obtained using the original samples. Then, we proceed as follows: for each year a random sub-sample of size 2,868 is drawn from the original data (the smallest recorded sample size over all survey waves that we use). Then, the opportunity tree and the resulting IOP are estimated and stored. This procedure is repeated 200 times. Doing so, the variables used, the splitting points, and the number of terminal nodes might differ not only across waves, but also across iterations for the same year. Finally, we report the average number of types and IOP values for each year.

Iterating the estimation of regression trees we get close to what in machine learning is known as *random forest*; a large collection of trees obtained using a subset of the available information. Contrary to the typically applied random forest procedure, we do not use a subset of regressors for each possible split of each tree. In contrast, the use of a subset of observations determine a certain level of heterogeneity of the tree structure across iterations. This reflects the level of uncertainty we have about the real data generating process, which, in this case, is our uncertainty about the real Romerian types existing in the German society. The level of IOP therefore becomes the average of 200 possible levels of IOP under alternative assumptions about the type partition.

5 The Evolution of IOP in Germany, 1992-2016

5.1 Development of the Opportunity Tree

Figure 2 shows the opportunity tree in 1992 and 2016. Both partitions are obtained using the full sample, and show a different structure of opportunities in Germany society for the two points in time. Appendix B shows the tree for all the other years over this time period.

[8] 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 scale back all incomes below the 0.1th percentile and exceeding the 99.9th percentile of the year-specific outcome distribution to these thresholds.

The just-unified Germany is a polarized society, with one main driver of between-type inequality: place of residence in 1989. For Germans originally from the East the second significant circumstance is father's occupation. Those having a father with a high occupation, or employed in the armed forces, show a higher level of expected outcome than those with a father in an unskilled occupation. Both East-types have an expected outcome below 80% of the national average for their age-group. For people residing in West-Germany, the splitting node is instead defined by Fathers' education. The average level of outcome in both western types is way above the national average.

The structure of opportunity is much more complex in 2016; almost the entire set of circumstances, excluding number of siblings and migration background, determine at least one splitting point. Interestingly, the place of residence in 1989 is still a fundamental driver of inequality and the second most correlated circumstance with the outcome. The first splitting point is determined by father's occupation; with armed forces occupations now together with unskilled occupations at the bottom of the distribution.

In 2016, the German society appears to be composed of a much larger number of types in comparison with the simpler opportunity structure suggested by the data for 1992. The vector of circumstances used in 1992 is location in 1989, father's occupation and education. In addition, for 2016 disability, sex, mother's education, and father's training play a role in splitting the sample. Furthermore, people who were resident abroad in 1989, and migrated to Germany afterwards, form the lower types together with individuals from the East. However, among these sub-types migration background does not contribute to further explaining the opportunity structure beyond other circumstances, like parental background and disability.

Figure 3 (and Table A6 in Appendix A) shows the development of the number of terminal nodes (types) from 1992 to 2016. The number of types gradually increases until the early 2000s with dramatic rises in 2000 and 2002, when the number jumps first to 13 and then to 20. Thereafter, the trend is characterized by ups and downs within this higher range.

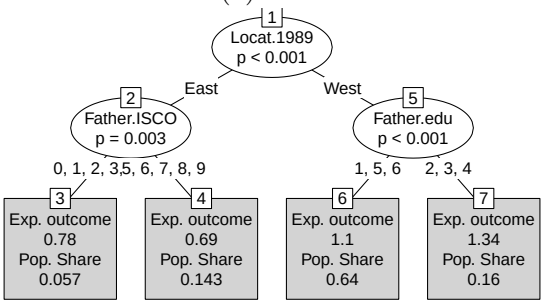
A closer look at the development of the opportunity tree from 1992 to 2016 reveals some striking patterns (see Appendix B). The first, most compelling evidence is that, albeit the appearance of some characteristics on the opportunity tree and the rising complexity of German society, the interaction of circumstances that defines the highest and the lowest type in the income distribution is rather constant. At the top of the distribution we always find individuals that resided in West-Germany before the fall of the Berlin Wall, whose parents had a high occupational position, and whose mothers had a high educational degree, while East-Germans with low educated parents persistently qualify at the lowest end.

Until 2001, location in 1989 is the first terminal node of the tree, and for the rest of the time it is a circumstance that persistently splits German society. To give an example, in 2014 people from East-Germany with high parental occupational background (Managers and Professionals) have lower average income levels than their West-German counterparts with middle parental occupation (Technicians and Clerks). Having a disability is, particularly from 2002 on, a circumstance that consistently explains inequality in Germany. This does not depend on the relative size of this group in the survey, since the share of respondents reporting disability oscillates without much variation around nine percent.

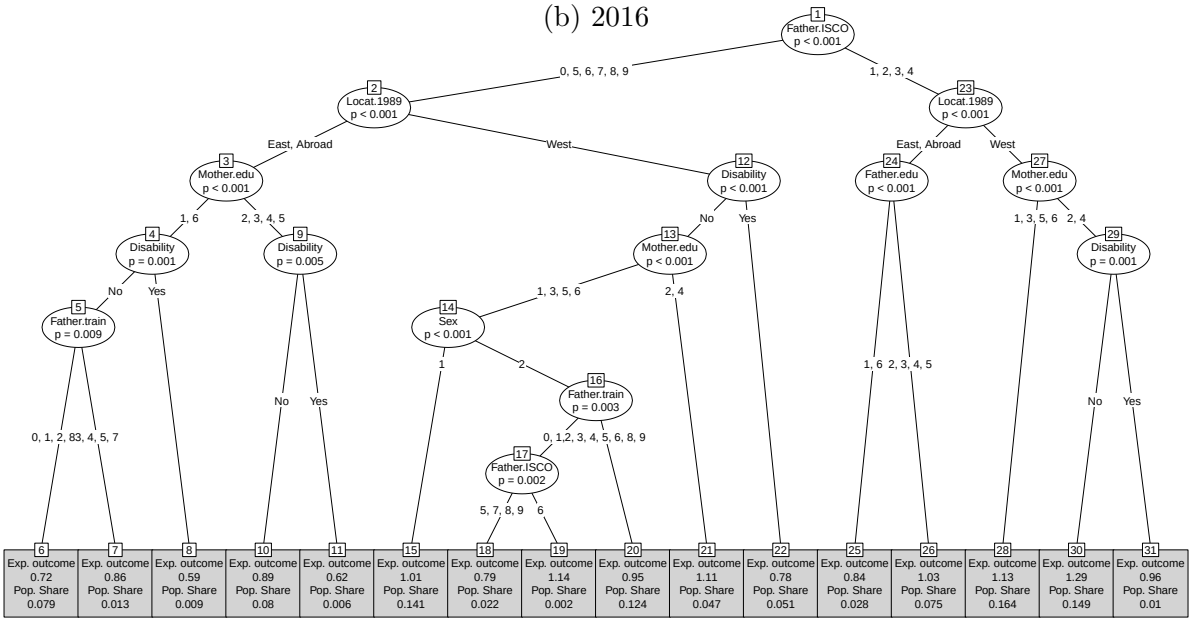
Characteristics like the month of birth, the predominant place in childhood (rural or urban), and the presence of siblings does not seem too relevant to explain the distribution of outcomes. Migration background appears relevant to the splitting of the West-German population in subtypes from 1999 to 2013. This applies to people with own migration experience (direct migration background) that moved to West-Germany before 1989. In contrast, the average income of the children of migrants (indirect migration background) is not distinguishable to the average income of the rest of the population; i.e. this circumstance mostly does not play a major role after having controlled for parental occupation and education, confirming past findings on the

Figure 2: Opportunity Tree in 1992 and 2016

(a) 1992

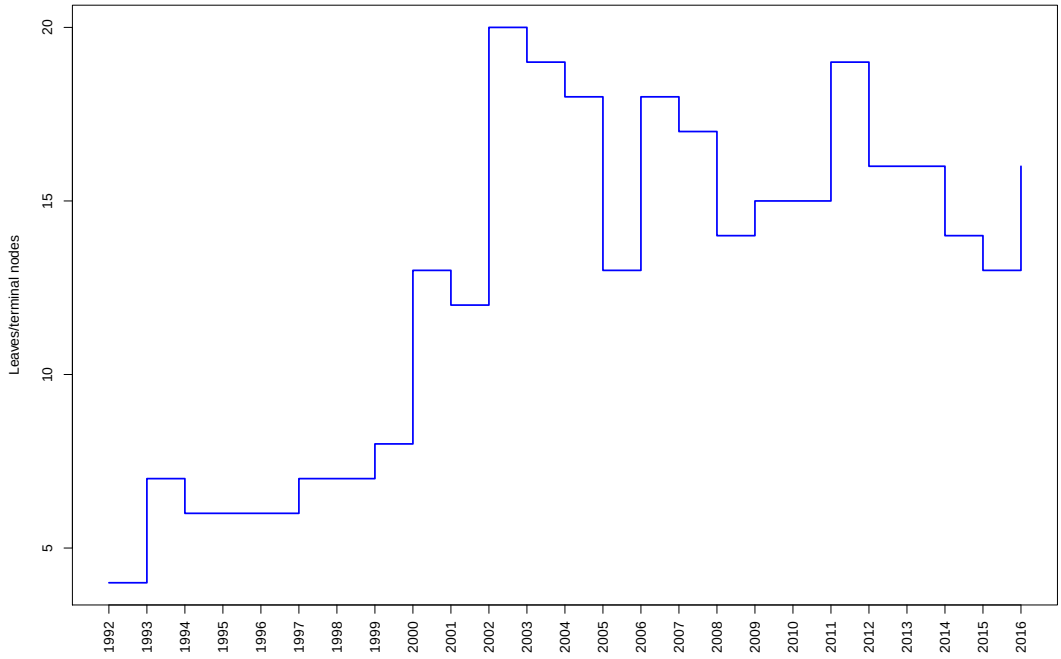


(b) 2016



Notes: Years refer to the survey wave. Incomes reported in the survey wave refer to the year before. **Father/Mother Education** : 1=Lower Secondary , 2=Intermediate Secondary, 3=Technical School, 4=Upper Secondary, 5=Other School Degree, 6=No School Degree, 7=School not attended **ISCO**: 0=Armed Forces, 1=Managers, 2=Professionals, 3=Technicians, 4=Clerks, 5=Service Workers, 6=Skilled Agricultural Workers, 7=Craftsmen, 8=Plant and Machine Workers, 9=Elementary Occupations. **Training**: 0=No information, 1=No vocational degree, 2=Vocational Degree, 3=Trade or Farming Apprentice, 4=Business, 5=Health Care or Special Technical School, 6=Civil Service Training, 7=Tech Engineer School, 8=College, University , 9 = Other Training. *Source: SOEPv33, 1992 and 2016.*

Figure 3: Number of types (terminal nodes) over time



Notes: Years refer to the survey wave. Incomes reported in the survey wave refer to the year before. *Source: SOEPv33, 1992-2016.*

topic (e.g. Bönke and Neidhöfer, 2018; Krause et al., 2015). New migrants, i.e. people that were resident Abroad in 1989, mostly belong to the lower types together with people that resided in East-Germany. However, only in one year, 2001, their average age-adjusted incomes are significantly different from the outcomes of East Germans. In this year, new migrants with low parental education form the type with the lowest age-adjusted income.

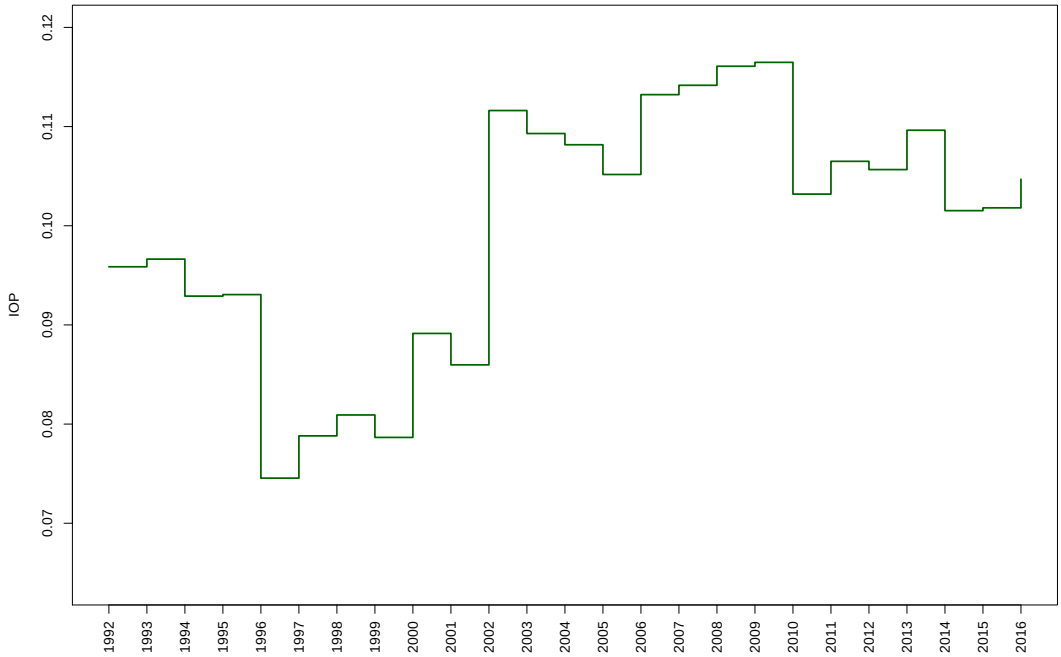
5.2 IOP estimates

Figure 4 (and Table A6 in Appendix A) shows the development of IOP, measured by the Gini coefficient, from 1992 to 2016. The IOP trend is characterized by declining inequality in the nineties and a subsequent rise, leading to a rather stable trend with little variation from 2003 onwards. Interestingly, despite the 2016 partition has four times the number of types than the 1992 partition (see Figure 3), the level of IOP is only slightly higher in 2016 (0.1046) than in 1992 (0.0959). Germany appears a much more complex society, with a complicated interaction of circumstances in producing opportunity. But the level of the resulting inequality is just slightly higher.

This becomes even more evident looking at how the type-specific empirical cumulative distribution functions (ECDFs) change over time. Figure 5 shows the type-specific ECDFs in 1992 and 2016; graphs for all the other years can be found in Appendix B. The dots are observed distributions while dashed lines show the interpolation of the distribution obtained applying the Bernstein polynomial approximation.

The comparison of the ECDFs in 1992 and 2016 is an illustrative example for the spectacular changing of IOP in Germany. In 1992, the two compressed distributions of East-residents in 1992 lie rather close to each other, while the more dispersed Western distributions lie far to the right. This polarization is no longer evident in 2016. The 16 type-specific distributions lie close

Figure 4: Development of IOP in Germany, 1992-2016



Notes: Years refer to the survey wave. Incomes reported in the survey wave refer to the year before. *Source: SOEPv33, 1992-2016.*

to each other and cross in several points. However, the distance between the highest and the lowest type is remarkable and remains stable over the entire period. This explains why the huge increase in the number of types has not been accompanied by a drastic rise in IOP.

5.3 Fixed Sample Size

As already discussed, opportunity trees and corresponding type-specific outcome distributions might not be strictly comparable due to changes in sample size. We therefore proceed by fixing the number of observations at 2,868 (the minimum sample size) and repeating the entire estimation procedure 200 times. Figure 6 shows the average number of terminal nodes and the level of IOP obtained by an iterative procedure with 200 repetitions. The reported bounds show the 0.975 and 0.025 quantiles of the distribution of the estimates.[9]

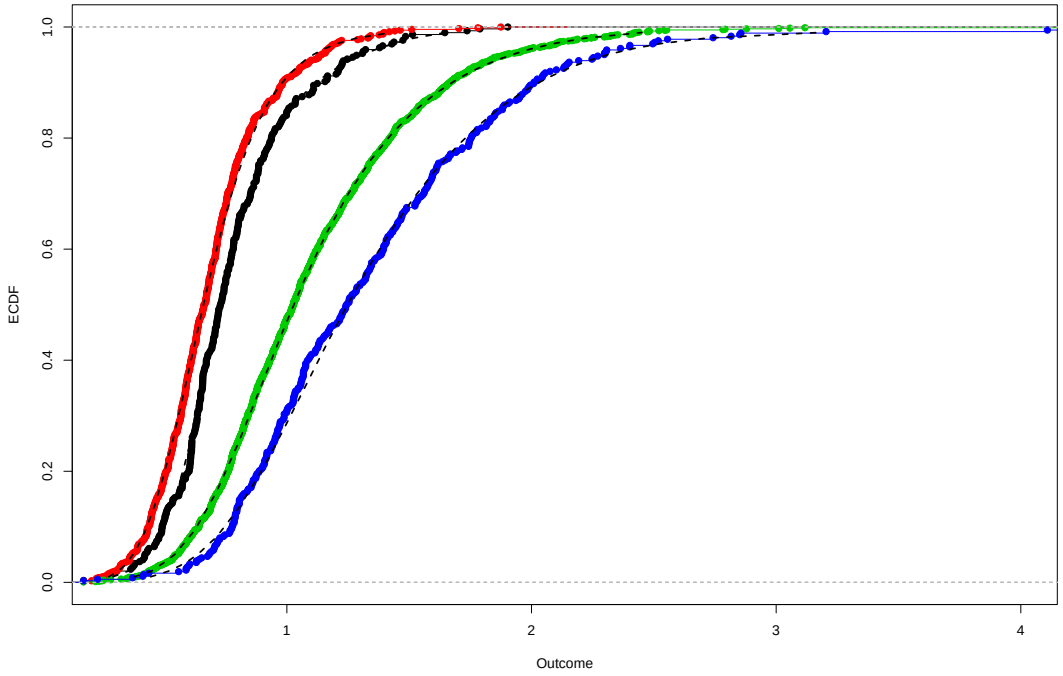
In comparison to the main analysis, in this sensitivity test the number of terminal nodes is reduced. The maximum number of types over the 200 iterations is on average 9.25 against the 20 obtained using the entire sample. Nevertheless, the trends in IOP, as well as in the number of types, are similar and show both an increase in the level of complexity over time, and an opportunity structure in 2016 markedly more complex than in 1992. The trend in IOP is also close to what is obtained with the full sample with a decline during the '90s and a steep increase in the early '00s. Again, the level of IOP in 2016 appears slightly above the level of 1992.

Estimating 200 trees with different samples makes it impossible to show opportunity trees for this application. Instead, for each circumstance we have a closer look at the share of trees in

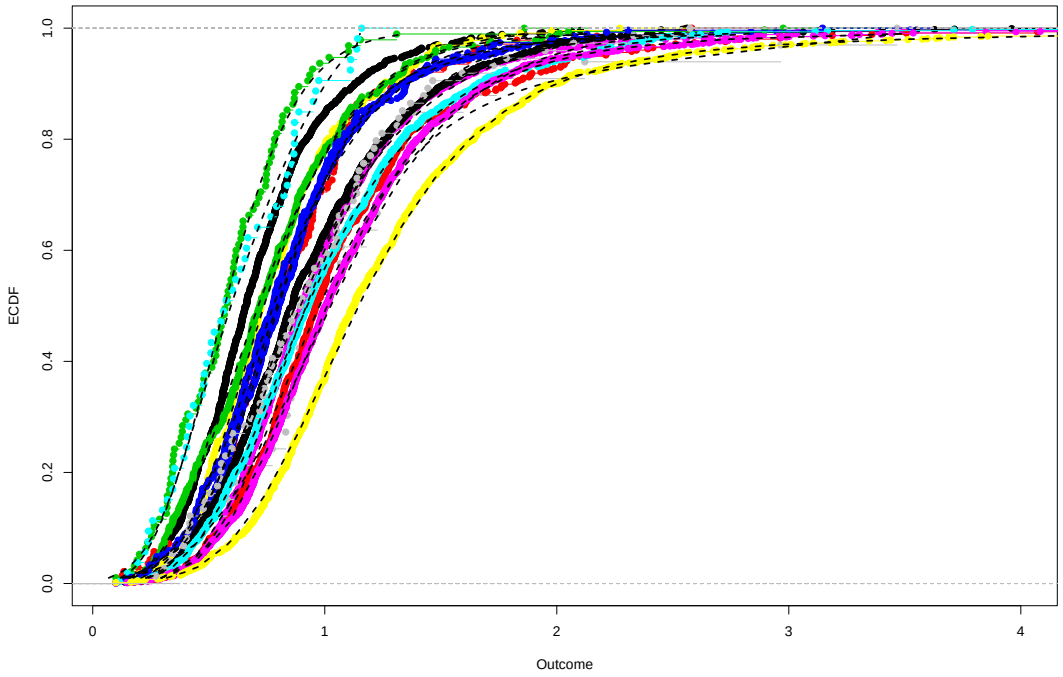
[9]Note, that bounds cannot be interpreted exactly as bootstrap confidence intervals: we do not resample with replacement, and we do not draw samples of the same sample size of the initial distribution. This also explains why there is no variability around the point estimates for 1992.

Figure 5: ECDFs in 1992 and 2016

(a) 1992



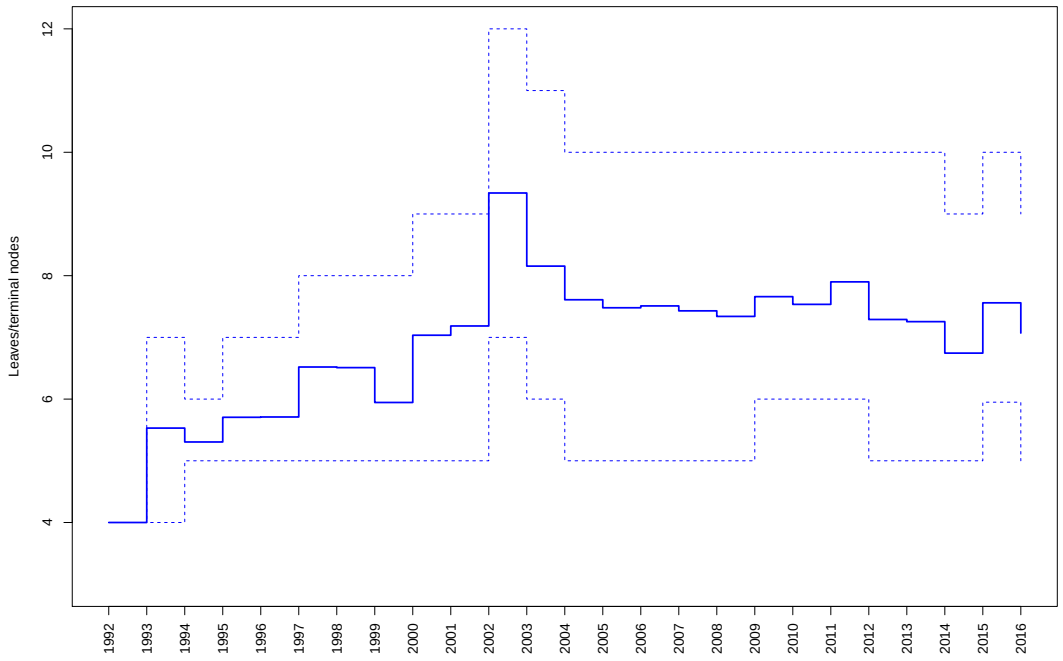
(b) 2016



Notes: Years refer to the survey wave. Incomes reported in the survey wave refer to the year before. *Source: SOEPv33, 1992 and 2016.*

Figure 6: Average number of types (top) and IOP (bottom) controlling for sample size

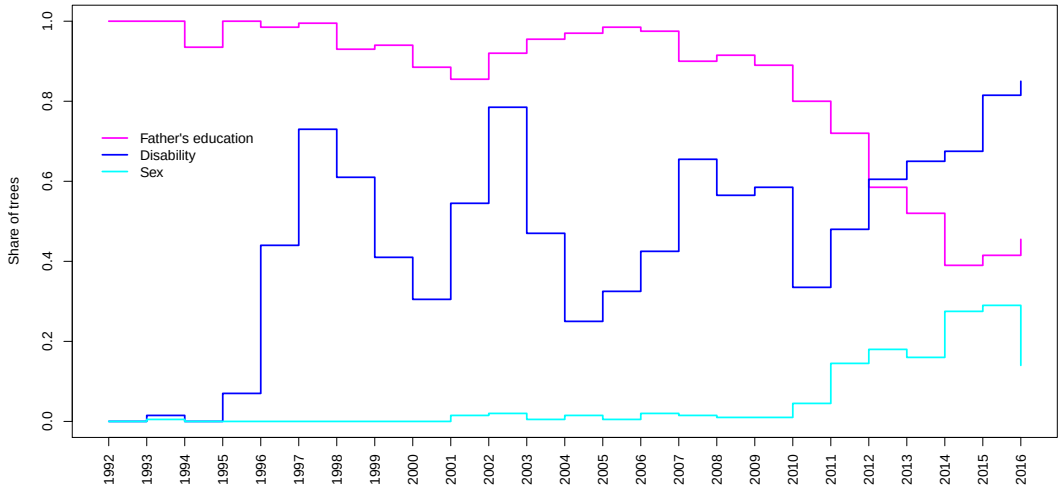
(a) Number of types



(b) IOP

Notes: Averages are calculated over 200 trees based on a sample of 2,868 observations drawn without replacement. Bounds show the 0.975 and 0.025 quantiles of the distribution of the estimates. *Source: SOEPv33, 1992-2016.*

Figure 7: Share of trees that use father's education, disability and sex to obtain Romerian types



Notes: Shares are calculated over 200 trees based on a sample of 2,868 observations drawn without replacement. *Source: SOEPv33, 1992-2016.*

which the variable determines at least one split (we report this shares in Table A5 in Appendix A). A striking finding is that location in 1989 is used in all 5,000 trees estimated for at least one split. Interesting trends can be observed for other circumstances. In Figure 7 we report how often father's education, disability and sex determine at least one split in the opportunity tree. The first circumstance shows a clear decline over time, while the second and the third, almost absent in the opportunity structure of the '90s, appear respectively in 80% and 20% of the trees in most recent waves. Hereby, the increasing role of sex in determining outcome inequality is likely to be explained by the increasing number of single parent households (from 13% in 1992 to 23% in 2016). In contrast, the increasing role of disability has no straightforward mechanical explanation, since the share of respondents reporting disability in SOEP is not growing over time (constantly between 7.5% and 10% from 1992 to 2016). The analysis of this remarkable pattern goes beyond the scope of this work, but should be addressed in future research.[10]

5.4 Discussion

Past studies report a rise in net income inequality in Germany in the 1990s until 2005/2006 and a subsequent stagnation characterized by small ups and downs (e.g. Biewen and Juhasz, 2012; Biewen et al., 2017; Jessen, 2019; Peichl et al., 2012). These studies identify changes in employment driven by part-time and marginal part-time work, and changes in the tax system as the major driver of this development, as well as the rising dispersion of labor market incomes due to skill-biased technological change (see Dustmann et al., 2009). Changes in the household size and structure, and reforms of the transfer system have been identified as minor influencing factors.[11]

The rising wage inequality in the nineties, evidenced e.g. by Fuchs-Schündeln et al. (2010), was not accompanied by rising IOP, as our analysis shows. Instead IOP drops, especially from

[10]The shares of single parents and disabled individuals is based on our own estimates using SOEPv33.

[11]Besides the economic literature on wage and income inequality in Germany, less attention has been dedicated to inequality of opportunity. Two exceptions are Peichl and Ungerer (2017) measuring East-West disparities in IOP, and Niehues and Peichl (2014) comparing IOP levels in Germany to the US.

1995 to 1996 and then slowly rises until 2001. Then, it experiences a sharp rise in 2002 that brings it to a new, higher level. Then IOP stays rather constant, with ups and downs, until 2016.[12] Part of the mechanisms described above could also explain first the decrease, then the sudden rise, and finally rather stagnant development of IOP in Germany from 1992 to 2016.

The first sharp increase is contemporaneous to major reforms of the tax and transfer system and of the unemployment benefit schemes. Especially the changes to the social benefit system also know as Hartz-reforms that were inacted in 2003, 2004, and 2005 had long lasting and controversially discussed effects on the German society. Past studies have shown an overall small income inequality-reducing effect of the reform, but with different impact on the middle and bottom of the distribution of income (Biewen and Juhasz, 2012). Particularly the incomes of longer-term unemployed were negatively affected, while social assistance receiver slightly gained from the reforms.

We do not observe a sizeable effect of the 2008-2009 financial crisis on IOP, confirming the conclusion of studies dedicated to income inequality. If any, we observe a downfall of IOP by 2 percentage points from 2009 to 2010. Bargain et al. (2017) find that in this period and until 2010 in Germany, policy changes induced a rise in poverty rates; mainly due to the slow adaption of social assistance, decreasing tax allowances and changes in the taxation of capital income. However, the tax reforms produced also lower marginal tax rates at both end of the distribution, and particularly for low levels of gross labor incomes the budget constraint rose from 2002 to 2011 (Jessen, 2019). This possible offset of mechanisms could explain, why the small rises and falls in IOP in this period are of minor magnitude. Another possible reason for this, is that several elements that explained the income inequality increase before 2005 became less strong over time. For instance, the rise in wage inequality became less steep, employment opportunities increased, and the middle and upper part of the distribution benefited from the employment boom after 2006 (Biewen et al., 2017).

6 Conclusions

We have suggested a novel approach to estimate IOP consistent with the theory proposed by Roemer (1998). In Roemer's view IOP is inequality due to circumstances beyond individual control. Outcome variability due to variables of choices is, instead, not part of IOP. The implementation of a measure consistent with this theory is complex because it necessitates both, to identify relevant circumstances beyond individual control, and to measure responsibility variables.

Our analysis borrows from machine learning methods and proposes a novel data-driven approach to the estimation of IOP. The main advantages of our approach are to minimize arbitrary assumptions about the composition of types and the shape of the type-specific outcome distributions. Romerian types, i.e. relevant interactions of circumstances, are obtained through conditional inference regression trees. The algorithm selects a partition in types that maximizes the outcome variability that can be consistently explained by between-type inequality. The identification of effort relies on a polynomial approximation of the empirical cumulative distribution function of outcomes in each type. The degree of the (Bernstein) polynomial is selected by a 10-fold cross-validation in order to maximize its out-of-sample log-likelihood.

We implement our method to 25 waves of the Socio-Economic Panel to describe the evolution of IOP in Germany between 1992 and 2016. We show that the structure of opportunities has markedly evolved over time. The partition in types detected by our approach in 2016 is much more complex than in 1992. We show that this difference is not driven by changes in the survey's structure; it persists also applying an iterative procedure that controls for changes in the sample

[12] Recall that the survey year refers to the incomes in the year before.

size.

Despite of the increase in the complexity describing the partition of the German society, the level of inequality of opportunity we measure in 2016 is only slightly higher than in 1992. The trend we observe sees a decline in IOP in the nineties, a sharp, sustained rise from 1999 to 2003 followed by another, smaller, sudden increase in 2006, and a subsequent stagnation with small ups and downs. The increase in IOP is suggestively contemporaneous with rising income inequality in the period 1999-2005 and the introduction of the Hartz-reforms. In contrast, a clear interconnection with the rising wage inequality in the nineties and the financial crisis is not visible. Mechanisms that could have offset these developments to cause a stronger role of circumstances to determine outcomes are a topic of great research interest for future studies.

Several further interesting suggestions arise from our analysis of the evolution of inequality of opportunity in Germany. The most compelling fact is that the East-West divide characterizes, more than two decades after reunification, most of the disparities in the German society. It is remarkable that even in recent times this circumstance divides the society in sub-types at the lower and at the higher end of the income distribution. Hence, over the entire observation period the types with the highest average level of outcome always consist of individuals that resided in West Germany in 1989. Besides, individuals with high educated mothers, often in combination with fathers in high occupational positions, are constantly part of the type with the highest level of outcome.

To sum up, our novel analysis of the development of IOP in Germany shows that, over the last two decades, the type with the highest level of outcome in the entire German income distribution consists of people with highly educated mothers and fathers in high occupational positions that resided in West-Germany before reunification. At the bottom of the distribution, we constantly find individuals with low educated fathers that resided in the former German Democratic Republic until the fall of the Berlin wall.

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.

Bargain, O., Callan, T., Doorley, K., and Keane, C. (2017). Changes in income distributions and the role of tax-benefit policy during the great recession: An international perspective. *Fiscal Studies*, 38(4):559–585.

Biewen, M. and Juhasz, A. (2012). Understanding rising income inequality in germany, 1999/2000-2005/2006. *Review of Income and Wealth*, 58(4):622–647.

Biewen, M., Ungerer, M., and Löffler, M. (2017). Why did income inequality in germany not increase further after 2005? *German Economic Review*, 0(0).

Bönke, T. and Neidhöfer, G. (2018). Parental background matters: Intergenerational mobility and assimilation of italian immigrants in germany. *German Economic Review*, 19(1):1–31.

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

Brunori, P., Hufe, P., and Mahler, D. G. (2018a). The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees. *ECINEQ Working Paper Series*, 2018-455.

Brunori, P., Peragine, V., and Serlenga, L. (2018b). Upward and downward bias when measuring inequality of opportunity. *Social Choice and Welfare*.

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

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

Dustmann, C., Ludsteck, J., and Schönberg, U. (2009). Revisiting the german wage structure. *The Quarterly Journal of Economics*, 124(2):843–881.

Dworkin, R. (1981). What is equality? part 2. equality of resources. *Philosophy and Public Affairs*, 10(4):283–345.

Farouki, R. T. (2012). The bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, 29:379–419.

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

Fuchs-Schündeln, N., Krueger, D., and Sommer, M. (2010). Inequality trends for germany in the last two decades: A tale of two countries. *Review of Economic Dynamics*, 13(1):103–132.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., and Schupp, J. (2019). The german socio-economic panel (soep). *Jahrbücher für Nationalökonomie und Statistik*, 239(2):345–360.

Hothorn, T. (2018). basefun: Infrastructure for computing with basis functions.

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.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Jessen, R. (2019). Why has income inequality in germany increased from 2002 to 2011? a behavioral microsimulation decomposition. *Review of Income and Wealth*, 0(0).

Krause, A., Rinne, U., and SchÃ¼ller, S. (2015). Kick it like Ãzil? decomposing the native-migrant education gap. *International Migration Review*, 49(3):757–789.

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.

Leblanc, A. (2012). On estimating distribution functions using bernstein polynomials. *Annals of the Institute of Statistical Mathematics*, 64(5):919–943.

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.

Luongo, P. (2011). The Implication of Partial Observability of Circumstances on the Measurement of Inequality of Opportunity. In Rodriguez, J. G., editor, *Inequality of Opportunity: Theory and Measurement*, volume 19 of *Research on Economic Inequality*, pages 23–49. Emerald, Bingley.

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.

Niehues, J. and Peichl, A. (2014). Upper bounds of inequality of opportunity: theory and evidence for germany and the us. *Social Choice and Welfare*, 43(1):73–99.

Peichl, A., Pestel, N., and Schneider, H. (2012). Does size matter? the impact of changes in household structure on income distribution in germany. *Review of Income and Wealth*, 58(1):118–141.

Peichl, A. and Ungerer, M. (2017). Equality of opportunity: East vs. west germany. *Bulletin of Economic Research*, 69(4):421–427.

Ramos, X. and va de Gaer, D. (2019). Is Inequality of Opportunity Robust to the Measurement Approach? *Review of Income and Wealth*, page forthcoming.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

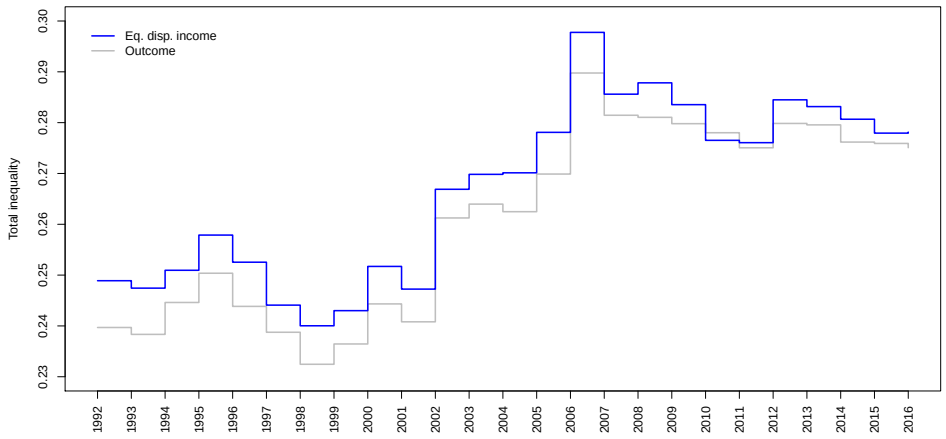
Rodríguez, J. G. (2008). Partial equality-of-opportunity orderings. *Social Choice and Welfare*, 31:435–456.

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

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

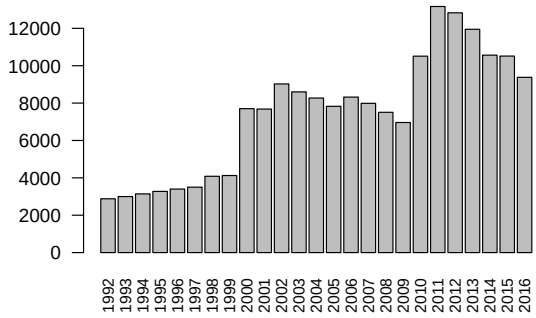
Appendix A Data and Estimates

Figure A1: Total inequality (Gini coefficient)



Notes: **Eq. disp. income** is the equivalized household disposable income. **Outcome** is the deviation of equivalent income from what expected given age (used in the main analysis). *Source: SOEPv33, 1992-2016.*

Figure A2: Number of observations per survey year



Source: SOEPv33, 1992-2016.

Table A1: Descriptive statistics

			Childhood in				Location in 1989		
Year	Female	Born in Ger.	Big town	Middle town	Small city	Countryside	East	West	Abroad
1992	0.51	0.95	0.24	0.17	0.21	0.38	0.20	0.80	0.00
1993	0.51	0.95	0.23	0.17	0.22	0.38	0.20	0.80	0.00
1994	0.50	0.95	0.24	0.17	0.22	0.38	0.20	0.80	0.00
1995	0.50	0.95	0.24	0.17	0.22	0.37	0.20	0.80	0.00
1996	0.50	0.95	0.24	0.17	0.22	0.36	0.19	0.80	0.01
1997	0.50	0.95	0.25	0.17	0.22	0.36	0.19	0.80	0.01
1998	0.50	0.95	0.24	0.17	0.23	0.36	0.19	0.80	0.01
1999	0.50	0.95	0.24	0.17	0.22	0.36	0.19	0.80	0.01
2000	0.50	0.92	0.23	0.18	0.22	0.36	0.20	0.77	0.03
2001	0.49	0.92	0.23	0.19	0.22	0.36	0.19	0.78	0.03
2002	0.49	0.91	0.22	0.19	0.22	0.37	0.19	0.77	0.04
2003	0.49	0.91	0.22	0.19	0.22	0.36	0.20	0.76	0.04
2004	0.49	0.90	0.22	0.19	0.22	0.36	0.19	0.77	0.04
2005	0.49	0.90	0.22	0.19	0.22	0.37	0.19	0.76	0.04
2006	0.49	0.90	0.22	0.19	0.23	0.36	0.20	0.75	0.05
2007	0.49	0.89	0.22	0.18	0.23	0.36	0.20	0.74	0.05
2008	0.49	0.89	0.22	0.18	0.23	0.36	0.21	0.74	0.05
2009	0.49	0.89	0.21	0.18	0.24	0.37	0.22	0.73	0.05
2010	0.49	0.88	0.22	0.18	0.24	0.36	0.22	0.71	0.07
2011	0.50	0.87	0.22	0.18	0.24	0.36	0.22	0.70	0.08
2012	0.50	0.86	0.23	0.18	0.24	0.36	0.22	0.70	0.09
2013	0.50	0.89	0.23	0.18	0.24	0.36	0.23	0.72	0.05
2014	0.50	0.89	0.23	0.17	0.25	0.35	0.23	0.72	0.05
2015	0.50	0.88	0.24	0.18	0.24	0.34	0.22	0.71	0.07
2016	0.49	0.88	0.23	0.18	0.24	0.35	0.22	0.71	0.07
Total	0.49	0.91	0.23	0.18	0.23	0.36	0.21	0.75	0.04

Source: SOEPv33, 1992-2016.

Table A2: Descriptive statistics, cnt.

	Migration background					Household				
Year	No	Direct	Indirect	Siblings	Disabled	Disp. income	N. of comp.	N. of Children	Age	Year of birth
1992	0.93	0.05	0.02	0.87	0.08	28818.07	2.99	0.79	43.69	1948.31
1993	0.93	0.05	0.02	0.87	0.08	30661.54	2.97	0.80	43.69	1949.31
1994	0.93	0.05	0.03	0.87	0.09	31832.58	2.94	0.80	43.65	1950.35
1995	0.92	0.05	0.03	0.87	0.09	31865.14	2.89	0.78	43.91	1951.09
1996	0.93	0.05	0.02	0.87	0.10	32048.61	2.86	0.76	44.01	1951.99
1997	0.92	0.05	0.03	0.87	0.10	32217.42	2.84	0.76	43.90	1953.10
1998	0.92	0.05	0.03	0.87	0.09	32405.06	2.86	0.80	43.86	1954.14
1999	0.92	0.05	0.03	0.87	0.09	33711.33	2.84	0.79	44.05	1954.95
2000	0.89	0.08	0.03	0.88	0.08	34086.65	2.82	0.77	44.19	1955.81
2001	0.89	0.08	0.03	0.88	0.08	34960.31	2.81	0.76	44.17	1956.83
2002	0.88	0.09	0.03	0.88	0.08	36103.56	2.80	0.76	44.32	1957.68
2003	0.88	0.09	0.03	0.88	0.08	37090.07	2.79	0.74	44.56	1958.44
2004	0.87	0.10	0.03	0.89	0.09	37070.00	2.78	0.71	44.95	1959.05
2005	0.87	0.10	0.03	0.89	0.08	38095.54	2.77	0.70	44.92	1960.08
2006	0.86	0.10	0.04	0.88	0.09	38797.80	2.72	0.67	45.21	1960.79
2007	0.85	0.11	0.04	0.88	0.09	39212.28	2.71	0.65	45.40	1961.60
2008	0.85	0.11	0.04	0.89	0.09	40358.91	2.67	0.61	45.54	1962.46
2009	0.85	0.11	0.05	0.88	0.09	40508.75	2.64	0.58	45.62	1963.38
2010	0.83	0.12	0.05	0.88	0.09	40713.78	2.62	0.57	45.67	1964.33
2011	0.82	0.13	0.05	0.88	0.09	41955.87	2.67	0.62	45.46	1965.54
2012	0.81	0.14	0.05	0.89	0.09	42693.66	2.67	0.62	45.54	1966.46
2013	0.85	0.11	0.05	0.88	0.10	43832.39	2.60	0.57	46.08	1966.92
2014	0.84	0.11	0.05	0.88	0.10	44767.95	2.62	0.59	46.11	1967.89
2015	0.84	0.12	0.05	0.88	0.10	45149.99	2.59	0.57	45.99	1969.01
2016	0.83	0.12	0.05	0.89	0.10	46408.61	2.60	0.58	46.11	1969.89
Total	0.87	0.09	0.04	0.88	0.09	38120.53	2.74	0.68	44.94	1960.11

Source: SOEPv33, 1992-2016.

Table A3: Categories for parental education, training and occupation

Circumstance	Values and Labels
Father/Mother Education	1 Lower Secondary 2 Intermediate Secondary 3 Technical School 4 Upper Secondary 5 Other School Degree 6 No School Degree 7 School not attended
Father/Mother ISCO	0 Armed Forces 1 Managers 2 Professionals 3 Technicians 4 Clerks 5 Service Workers 6 Skilled Agricultural Workers 7 Craftsmen 8 Plant and Machine Workers 9 Elementary Occupations
Father/Mother Training	0 No information 1 No vocational degree 2 Vocational Degree 3 Trade or Farming Apprentice 4 Business 5 Health Care or Special Technical School 6 Civil Service Training 7 Tech Engineer School 8 College, University 9 Other Training

Source: SOEPv33.

Table A4: Share of individuals with missing information on circumstances

	Month of birth	East/West in 1989	Location in childhood	Disabled	Siblings	Born in Germany
1992	0.42	0.01	0.07	0.10	0.44	0.00
1993	0.40	0.01	0.06	0.10	0.42	0.00
1994	0.37	0.02	0.04	0.10	0.39	0.00
1995	0.35	0.03	0.04	0.10	0.37	0.00
1996	0.33	0.03	0.04	0.10	0.35	0.00
1997	0.30	0.04	0.04	0.10	0.32	0.00
1998	0.28	0.06	0.06	0.11	0.31	0.00
1999	0.25	0.07	0.07	0.12	0.28	0.00
2000	0.24	0.12	0.13	0.09	0.28	0.00
2001	0.22	0.11	0.13	0.16	0.26	0.00
2002	0.13	0.09	0.11	0.14	0.21	0.00
2003	0.08	0.04	0.08	0.14	0.13	0.00
2004	0.05	0.03	0.08	0.13	0.08	0.00
2005	0.05	0.02	0.08	0.13	0.05	0.00
2006	0.07	0.05	0.10	0.14	0.07	0.00
2007	0.07	0.05	0.10	0.16	0.07	0.00
2008	0.05	0.04	0.08	0.16	0.06	0.00
2009	0.08	0.07	0.11	0.16	0.09	0.00
2010	0.07	0.11	0.15	0.15	0.13	0.00
2011	0.04	0.07	0.11	0.16	0.10	0.00
2012	0.04	0.06	0.09	0.19	0.08	0.00
2013	0.06	0.07	0.08	0.20	0.08	0.00
2014	0.06	0.06	0.08	0.25	0.07	0.00
2015	0.06	0.07	0.08	0.24	0.08	0.00
2016	0.08	0.21	0.22	0.37	0.11	0.00
Total	0.13	0.07	0.10	0.17	0.16	0.00

	Migration Background	Father's education	Mother's education	Father's training	Mother's training	Father's ISCO-code
1992	0.00	0.29	0.28	0.29	0.30	0.40
1993	0.00	0.27	0.26	0.28	0.28	0.37
1994	0.00	0.28	0.27	0.30	0.30	0.38
1995	0.00	0.31	0.29	0.33	0.33	0.40
1996	0.00	0.30	0.29	0.32	0.32	0.38
1997	0.00	0.30	0.28	0.31	0.31	0.37
1998	0.00	0.28	0.26	0.30	0.30	0.36
1999	0.00	0.27	0.25	0.29	0.29	0.35
2000	0.00	0.27	0.25	0.28	0.29	0.33
2001	0.00	0.26	0.25	0.28	0.29	0.32
2002	0.00	0.23	0.21	0.24	0.25	0.28
2003	0.00	0.19	0.18	0.20	0.21	0.24
2004	0.00	0.18	0.17	0.20	0.20	0.24
2005	0.00	0.18	0.16	0.19	0.20	0.23
2006	0.00	0.19	0.18	0.20	0.21	0.23
2007	0.00	0.19	0.17	0.20	0.21	0.23
2008	0.00	0.17	0.16	0.18	0.19	0.21
2009	0.00	0.19	0.18	0.20	0.21	0.23
2010	0.00	0.23	0.21	0.23	0.23	0.26
2011	0.00	0.19	0.17	0.19	0.18	0.22
2012	0.00	0.17	0.15	0.16	0.16	0.20
2013	0.00	0.29	0.26	0.16	0.15	0.24
2014	0.00	0.30	0.27	0.15	0.14	0.24
2015	0.00	0.29	0.27	0.15	0.14	0.22
2016	0.00	0.39	0.37	0.28	0.27	0.33
Total	0.00	0.25	0.23	0.22	0.22	0.28

Source: SOEPv33, 1992-2016.

Table A5: Share of trees in which each circumstance determines at least one split

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Mig. background	0	0	0	0	0.17	0.165	0.16	0.405	0.615	0.46	0.41	0.405	0.49	0.28	0.125	0.13	0.365	0.345	0.6	0.305	0.08	0.165	0.1	0.14	0.08
Locat. 1989	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Mother edu	0	0.925	0.755	0.645	0.47	0.44	0.67	0.39	0.65	0.81	0.495	0.375	0.495	0.48	0.67	0.66	0.45	0.745	0.76	0.815	0.735	0.795	0.475	0.515	0.75
Father edu	1	1	0.935	1	0.985	0.995	0.93	0.94	0.885	0.855	0.92	0.955	0.97	0.985	0.975	0.9	0.915	0.89	0.8	0.72	0.585	0.52	0.39	0.415	0.455
Father ISCO	1	0.07	0.945	0.77	0.135	0.29	0.77	0.95	0.865	0.795	0.72	0.565	0.545	0.615	0.48	0.75	0.855	0.84	0.82	0.92	0.945	0.98	1	0.98	0.955
Father train	0	0.81	0.375	0.745	0.76	0.835	0.84	0.435	0.25	0.2	0.915	0.99	0.915	0.91	0.88	0.585	0.435	0.39	0.25	0.385	0.195	0.23	0.135	0.595	0.165
Sibling	0	0	0	0.005	0	0	0.015	0.135	0.1	0.085	0.275	0.2	0.14	0.135	0.19	0.24	0.16	0.075	0.17	0.165	0.065	0.125	0.095	0.075	0.035
Disability	0	0.015	0	0.07	0.44	0.73	0.61	0.41	0.305	0.545	0.785	0.47	0.25	0.325	0.425	0.655	0.565	0.585	0.335	0.48	0.605	0.65	0.675	0.815	0.85
Sex	0	0.005	0	0	0	0	0	0	0	0.015	0.02	0.005	0.015	0.005	0.02	0.015	0.01	0.01	0.045	0.145	0.18	0.16	0.275	0.29	0.14

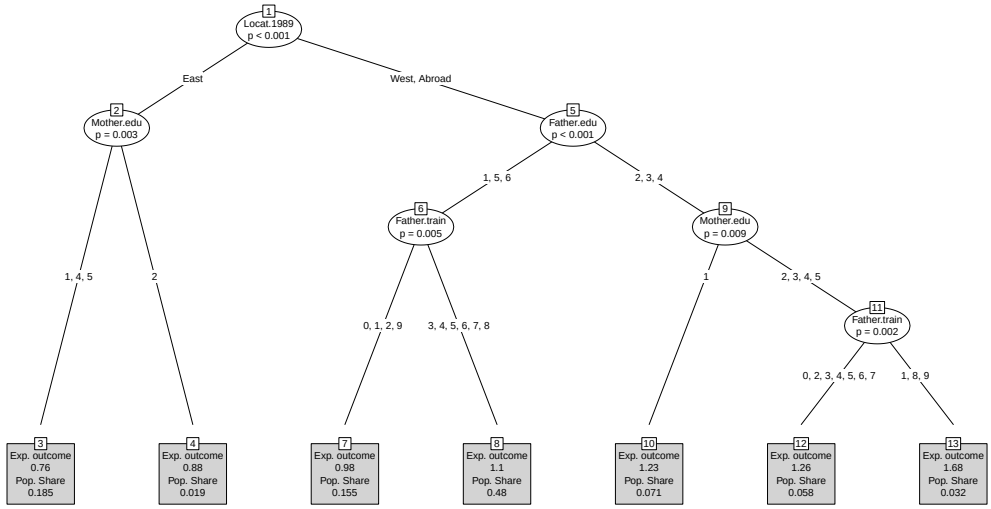
Notes: Shares are calculated over 200 trees based on a sample of 2,868 observations drawn without replacement. *Source: SOEPv33, 1992-2016.*

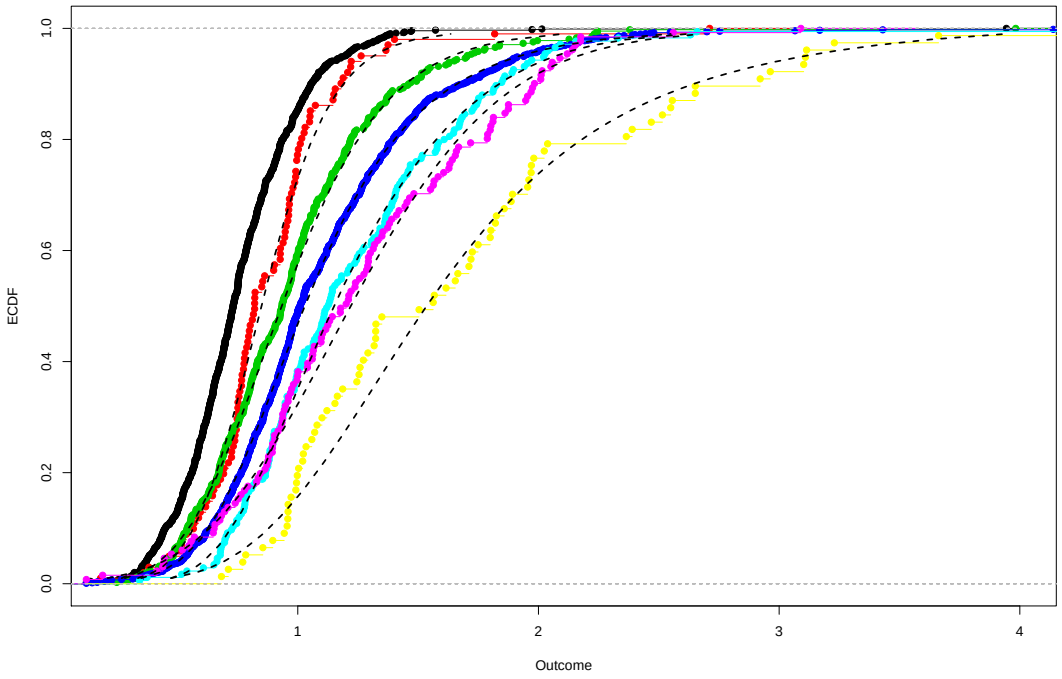
Table A6: Estimates

	Full sample		Same sample size					
Year	n. nodes	IOP	n. nodes	low	high	IOP	low	high
1992	4	0.0952	4	4	4	0.0952	0.0952	0.0952
1993	7	0.0961	5.53	4	7	0.0928	0.0874	0.0972
1994	6	0.0926	5.305	5	6	0.0889	0.0843	0.0939
1995	6	0.0932	5.705	5	7	0.0911	0.0859	0.0949
1996	6	0.0744	5.71	5	7	0.0762	0.0705	0.0812
1997	7	0.0779	6.52	5	8	0.0777	0.0727	0.0826
1998	7	0.0801	6.51	5	8	0.0787	0.0719	0.0846
1999	8	0.0784	5.945	5	8	0.076	0.0706	0.081
2000	13	0.0887	7.035	5	9	0.0835	0.0765	0.0902
2001	10	0.0847	7.185	5	9	0.0826	0.0739	0.0894
2002	20	0.1108	9.34	7	12	0.109	0.1007	0.1194
2003	19	0.1089	8.155	6	11	0.105	0.0967	0.115
2004	18	0.1081	7.61	5	10	0.1013	0.0926	0.1101
2005	13	0.1033	7.48	5	10	0.1034	0.0955	0.1124
2006	18	0.1135	7.51	5	10	0.112	0.1004	0.126
2007	17	0.1140	7.43	5	10	0.1094	0.0981	0.1203
2008	14	0.1164	7.34	5	10	0.1069	0.0963	0.1174
2009	15	0.1152	7.66	6	10	0.1059	0.0947	0.1186
2010	15	0.1031	7.535	6	10	0.0981	0.0881	0.1091
2011	19	0.1067	7.9	6	10	0.1	0.0893	0.1114
2012	16	0.1061	7.29	5	10	0.1006	0.088	0.1126
2013	16	0.1098	7.255	5	10	0.0964	0.0844	0.1083
2014	14	0.1013	6.745	5	9	0.0955	0.0843	0.1055
2015	13	0.1020	7.56	5.95	10	0.0987	0.0868	0.1095
2016	16	0.1046	7.07	5	9	0.0993	0.0895	0.1101

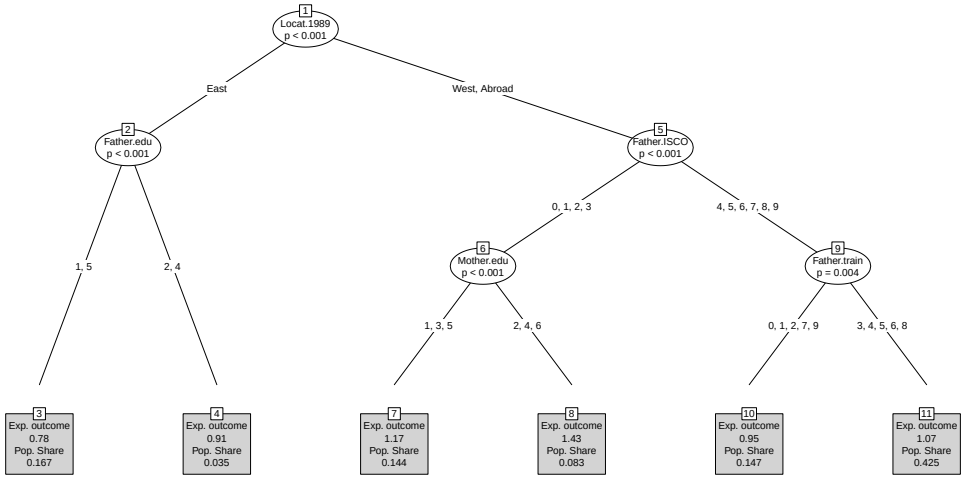
Notes: Values for identical sample size are calculated over 200 trees based on a sample of 2,868 observations drawn without replacement. *Source: SOEPv33, 1992-2016.*

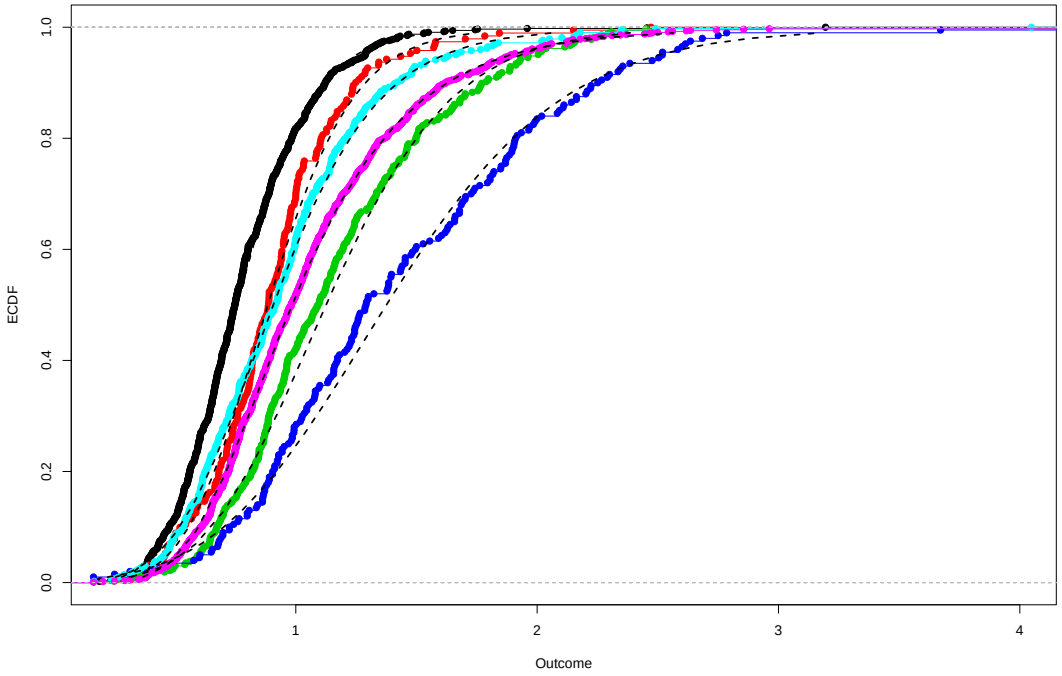
Appendix B Opportunity Tree and CDFs for single years



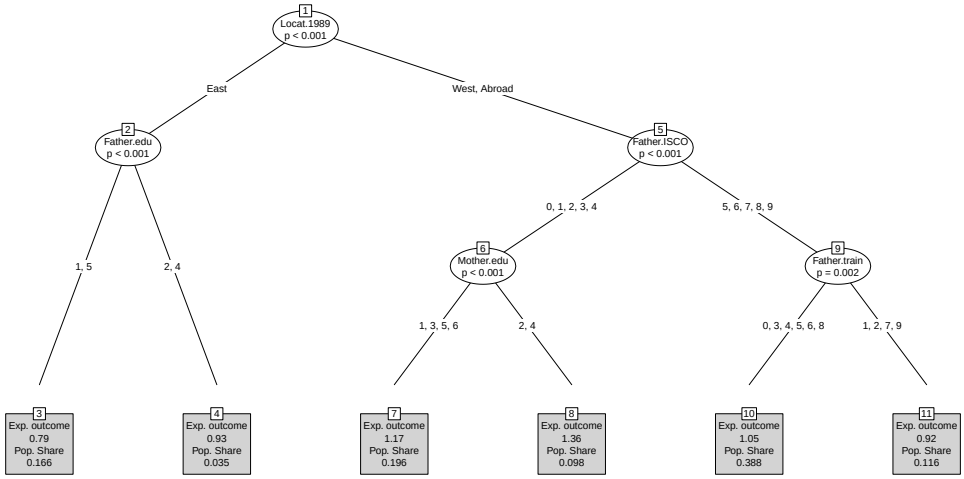


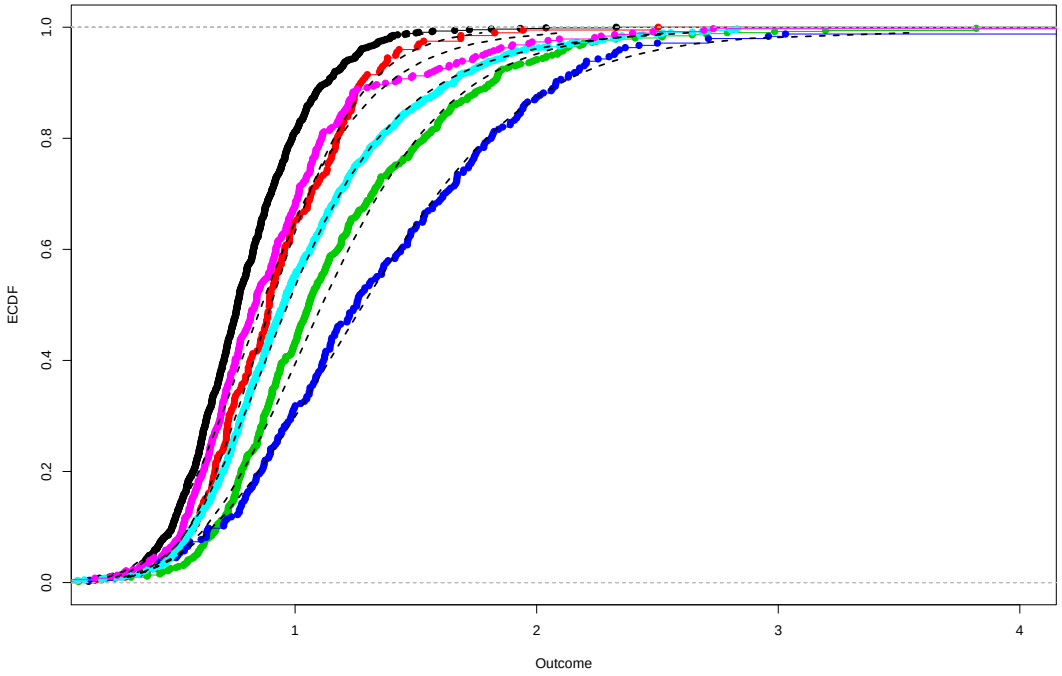
SOEPv33, 1993.



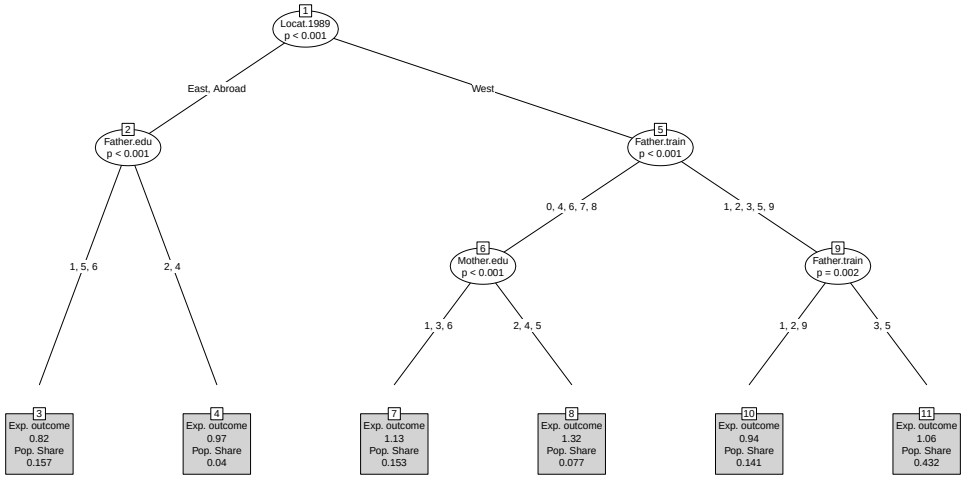


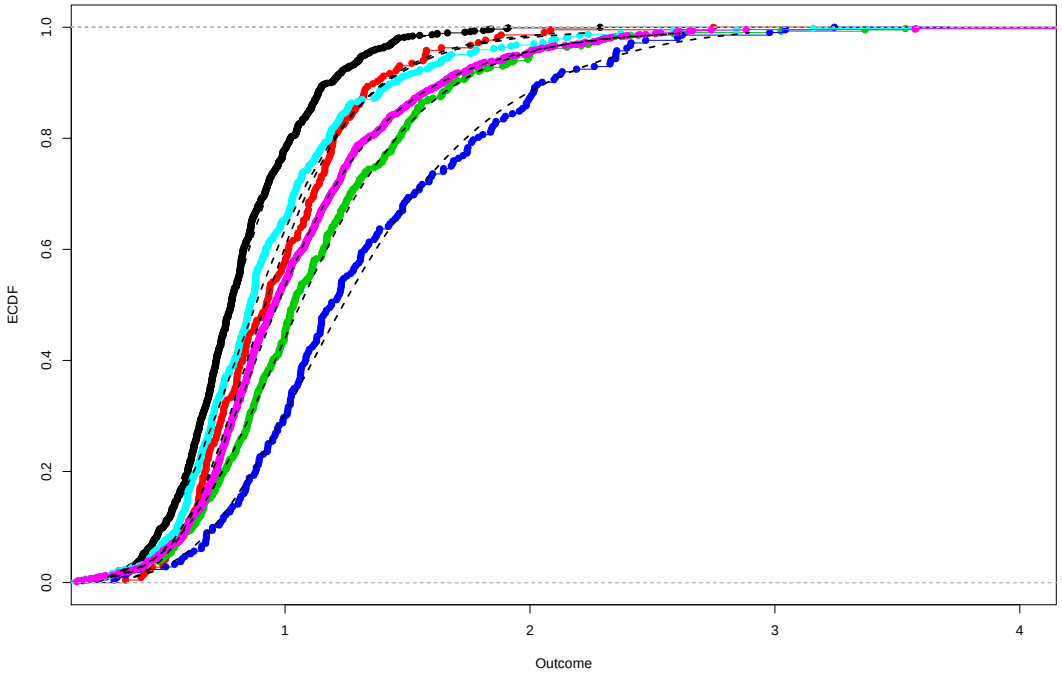
SOEPv33, 1994.



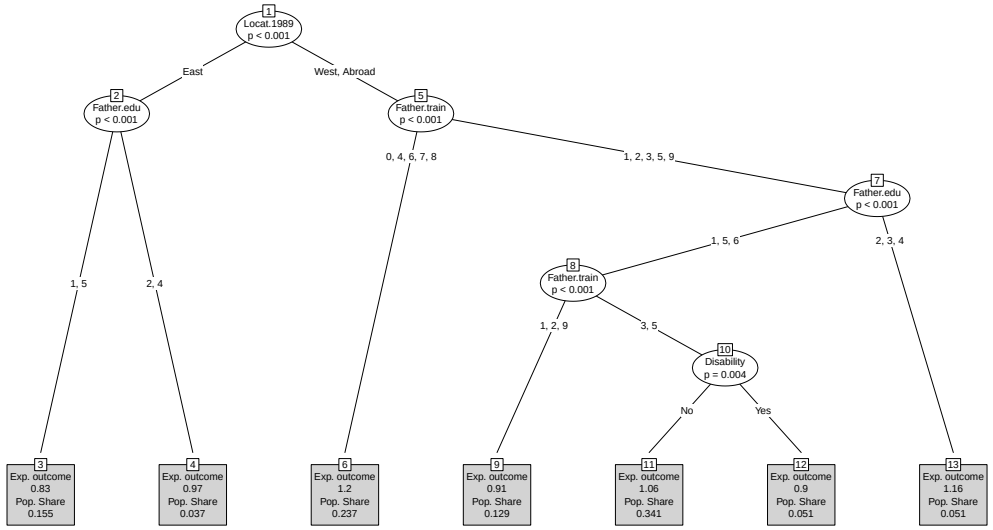


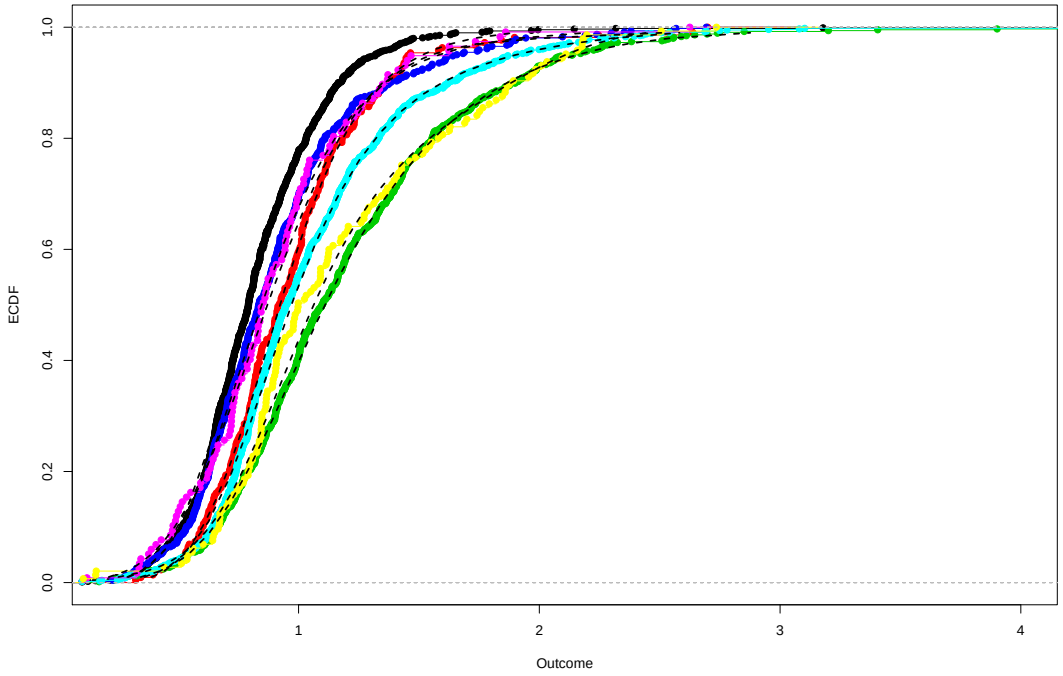
SOEPv33, 1995.



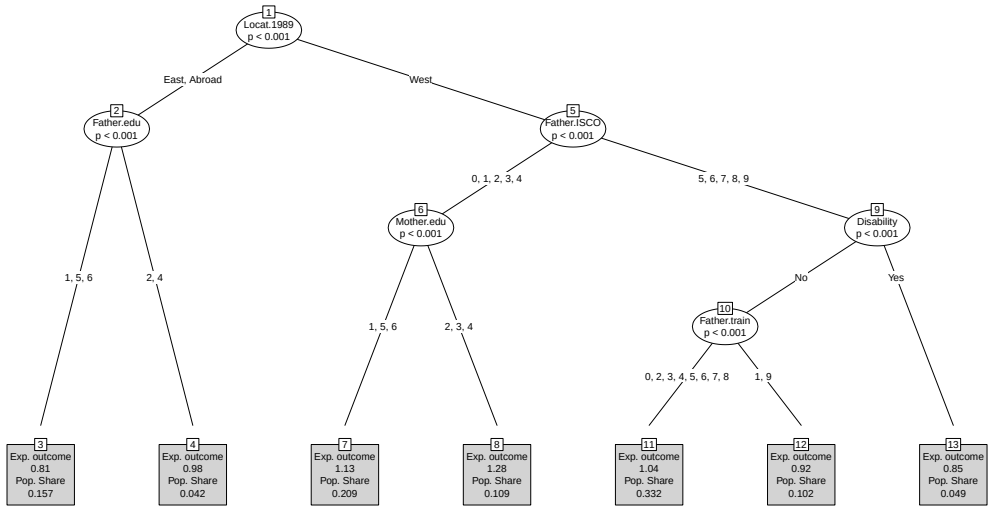


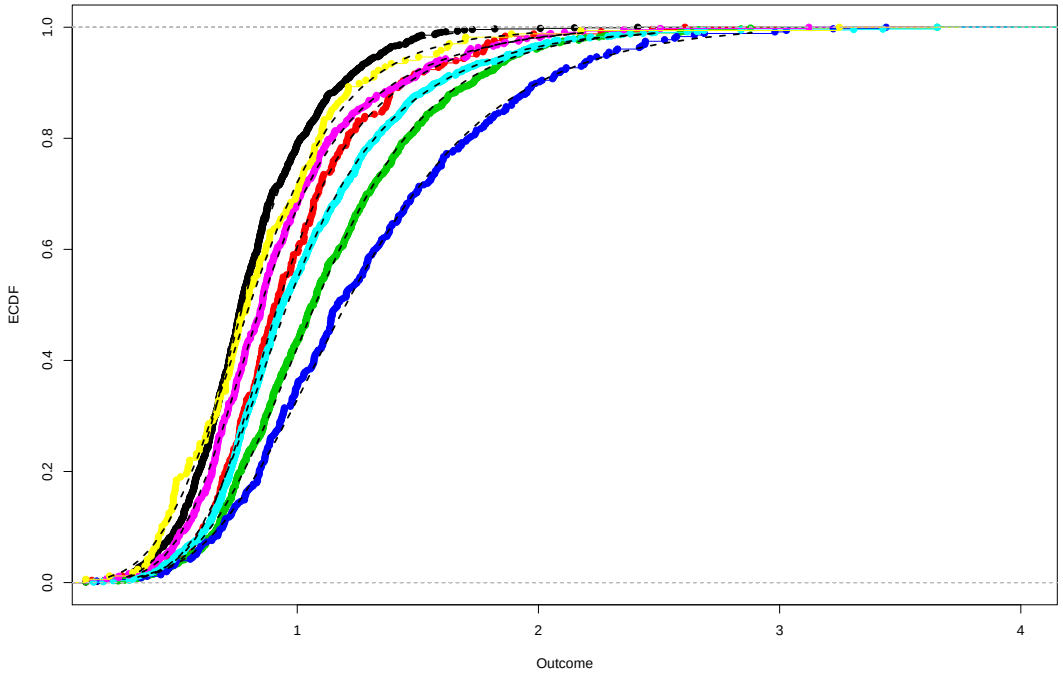
SOEPv33, 1996.



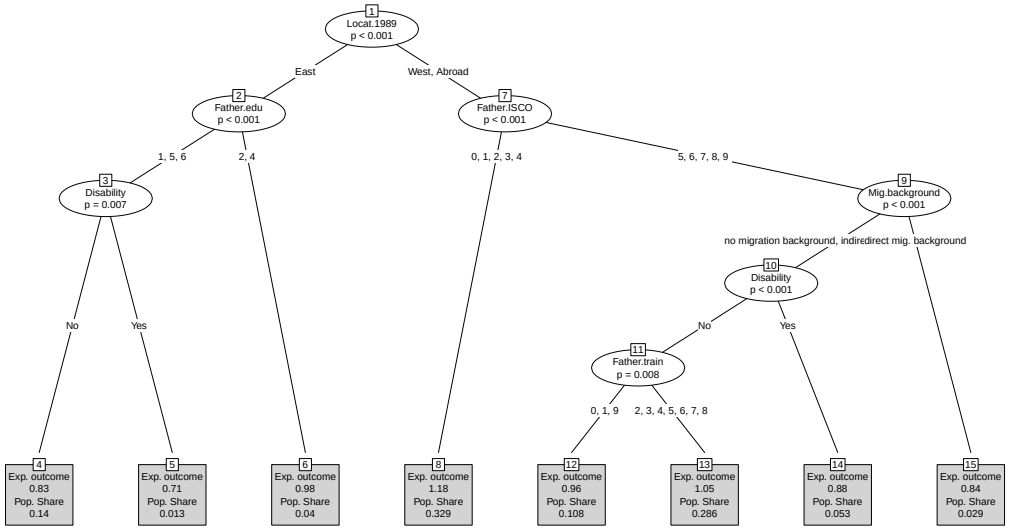


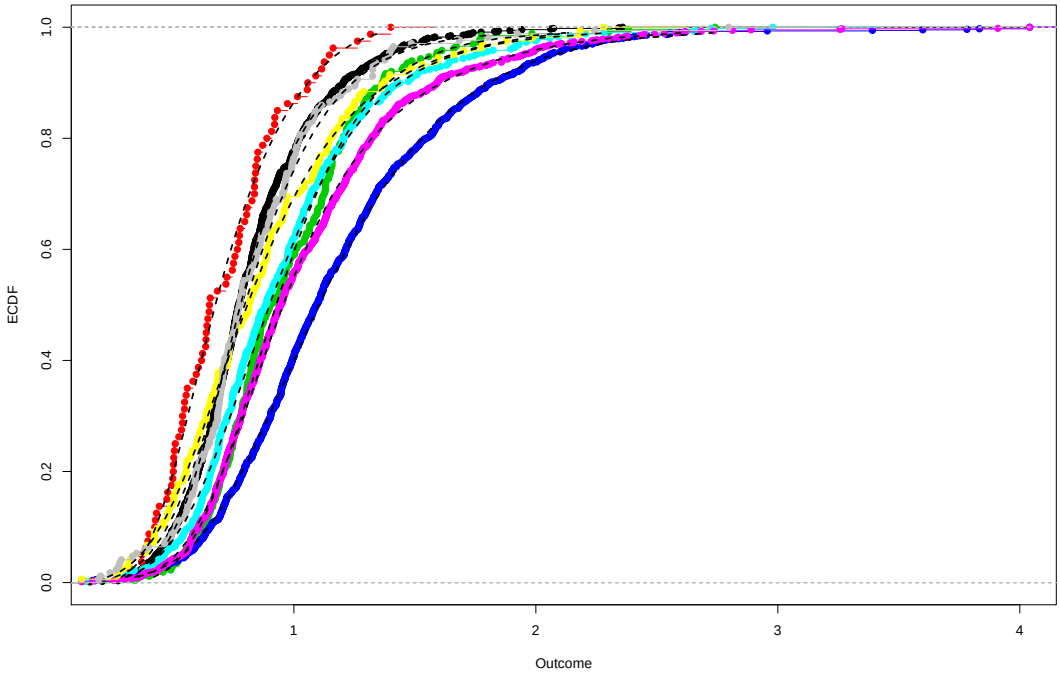
SOEPv33, 1997.



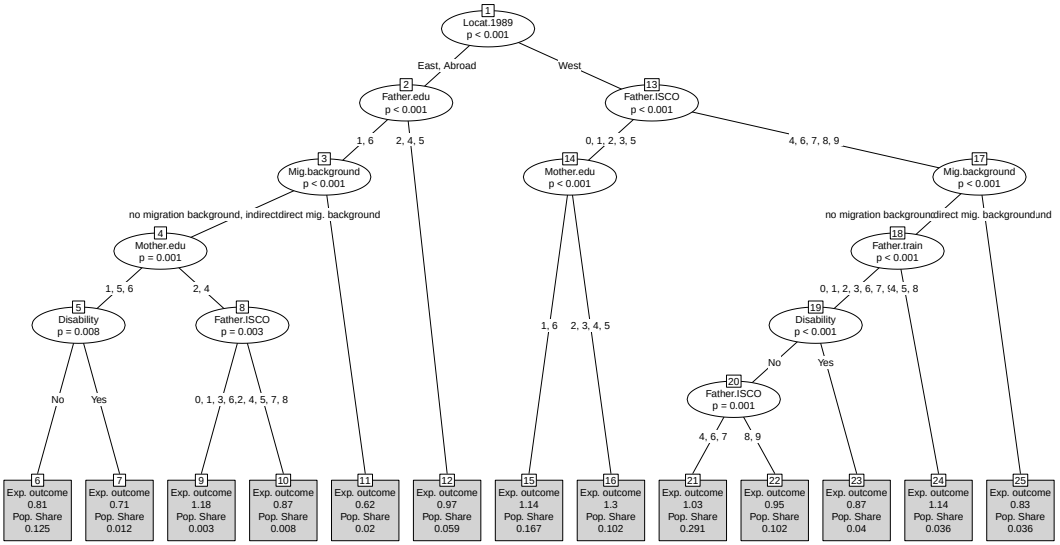


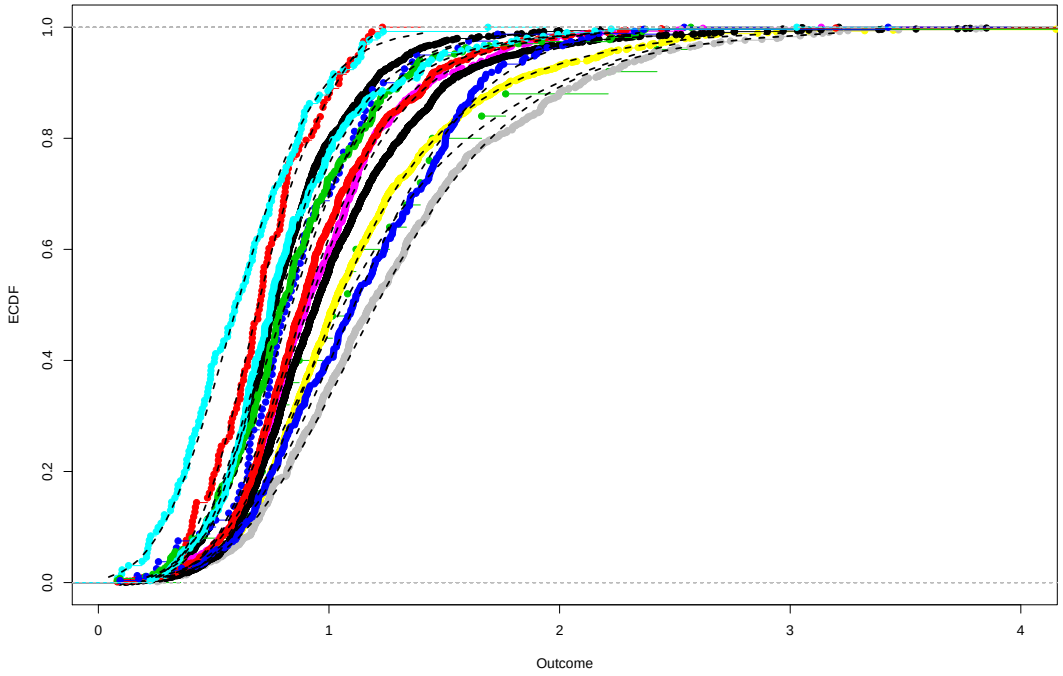
SOEPv33, 1998.



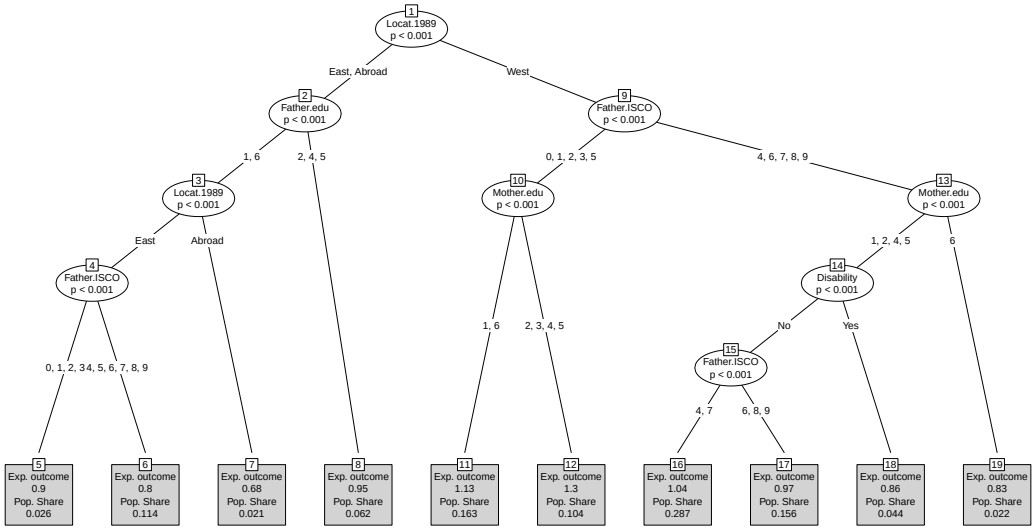


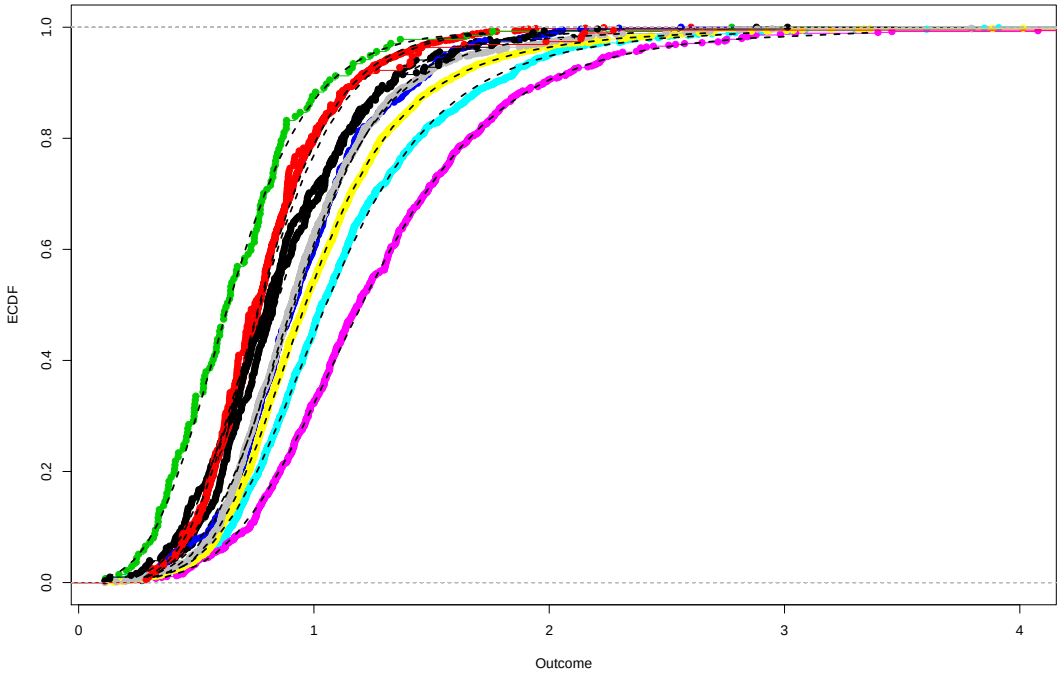
SOEPv33, 1999.



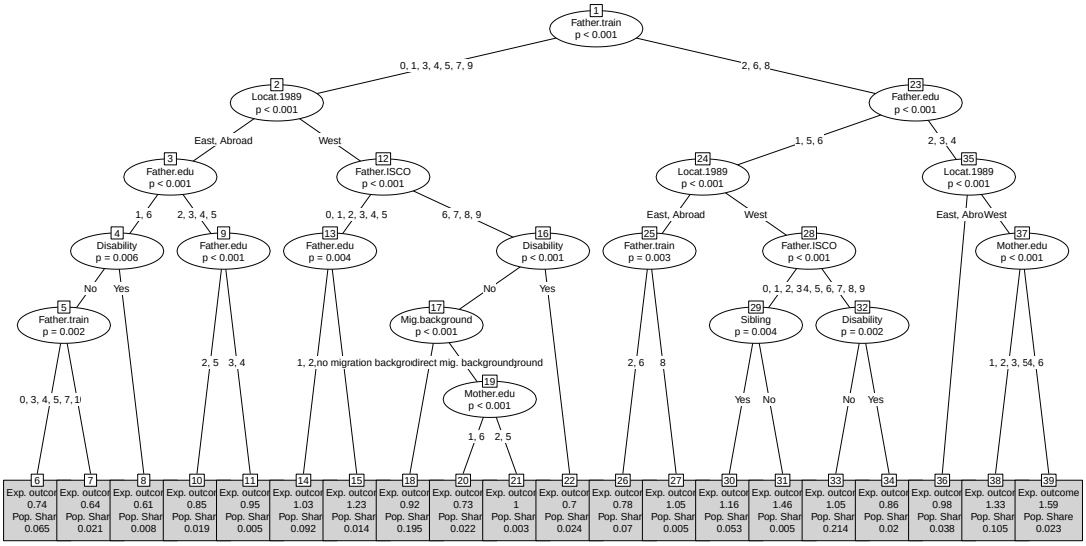


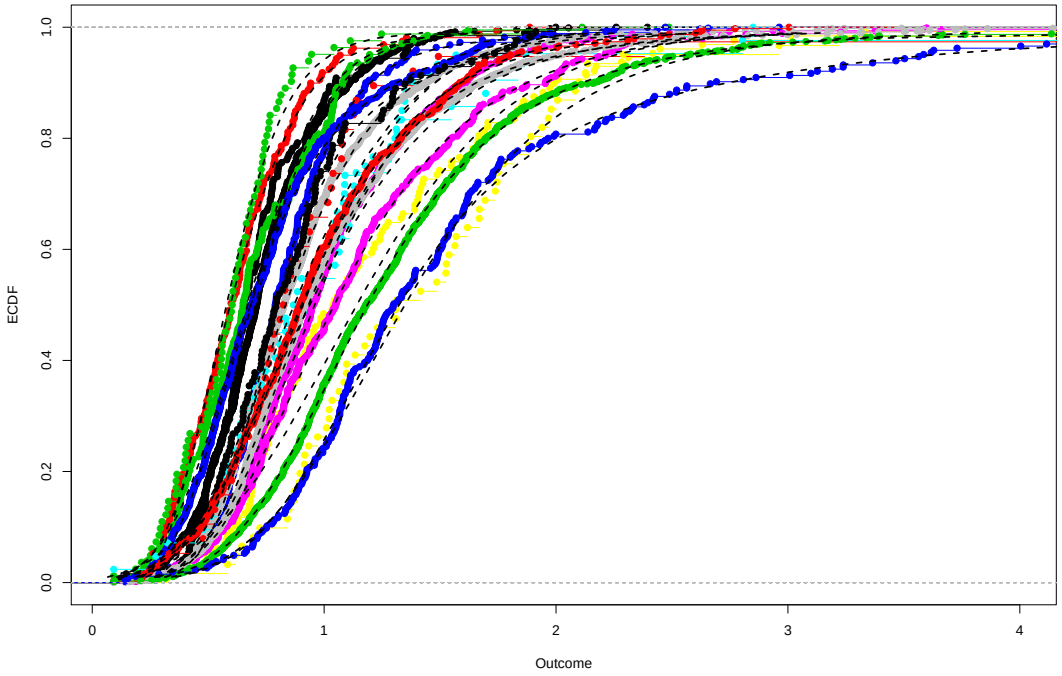
SOEPv33, 2000.



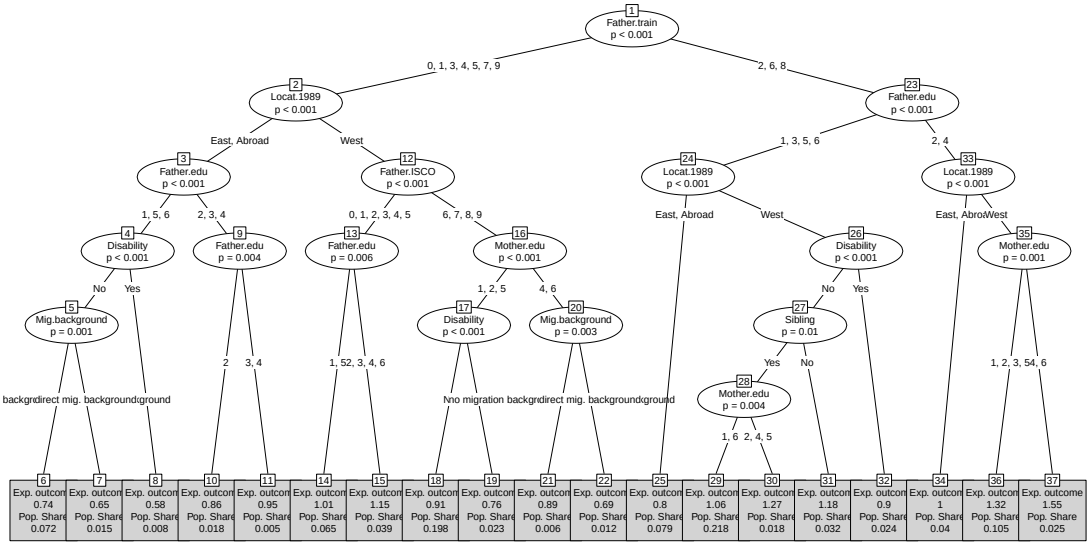


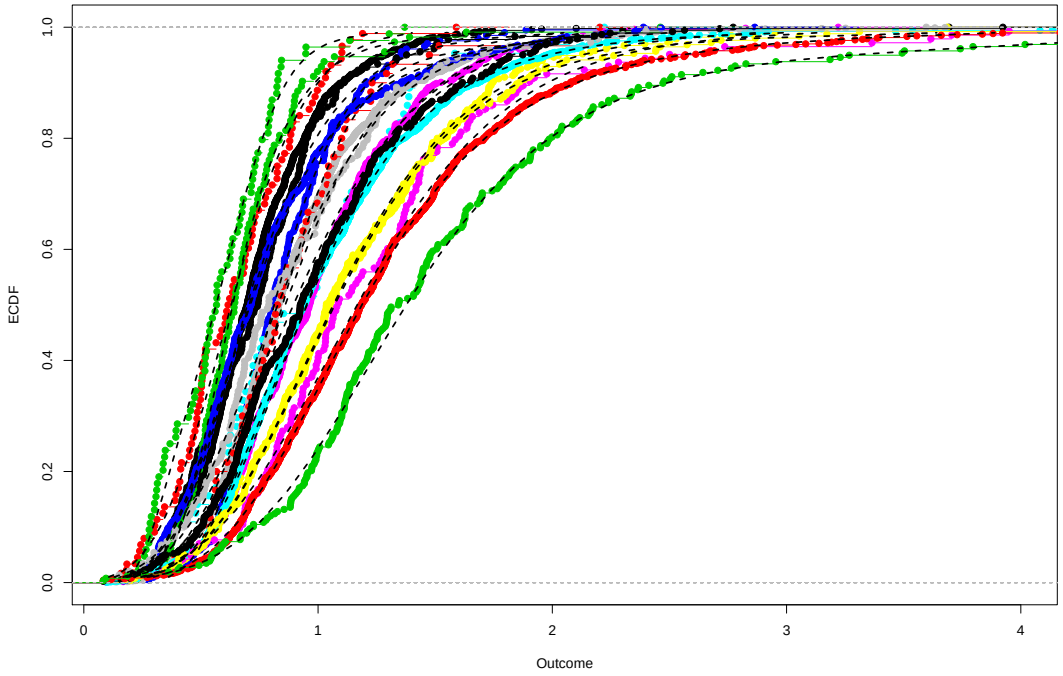
SOEPv33, 2001.



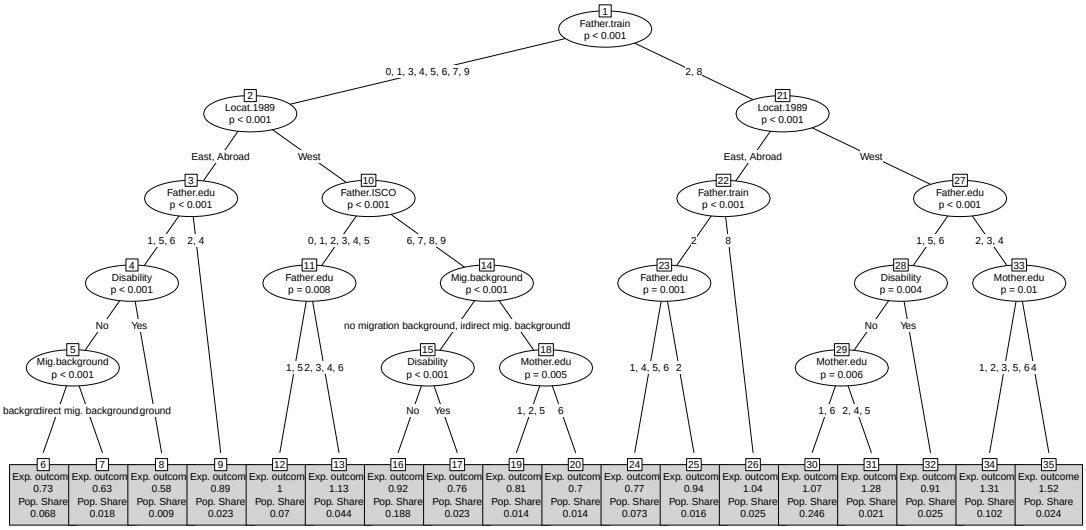


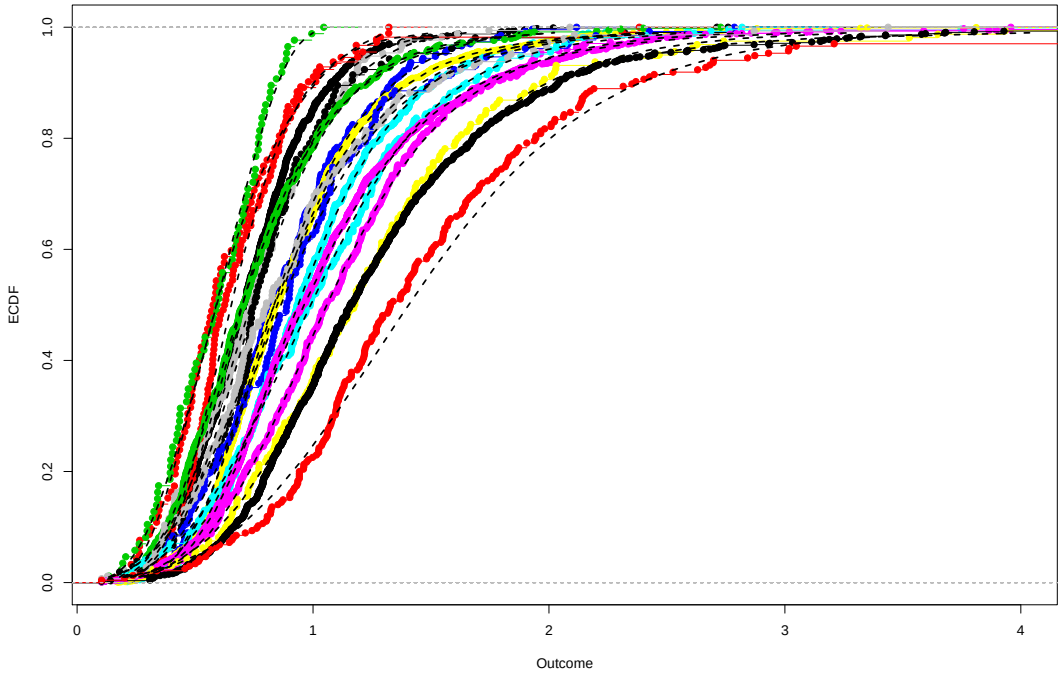
SOEPv33, 2002.



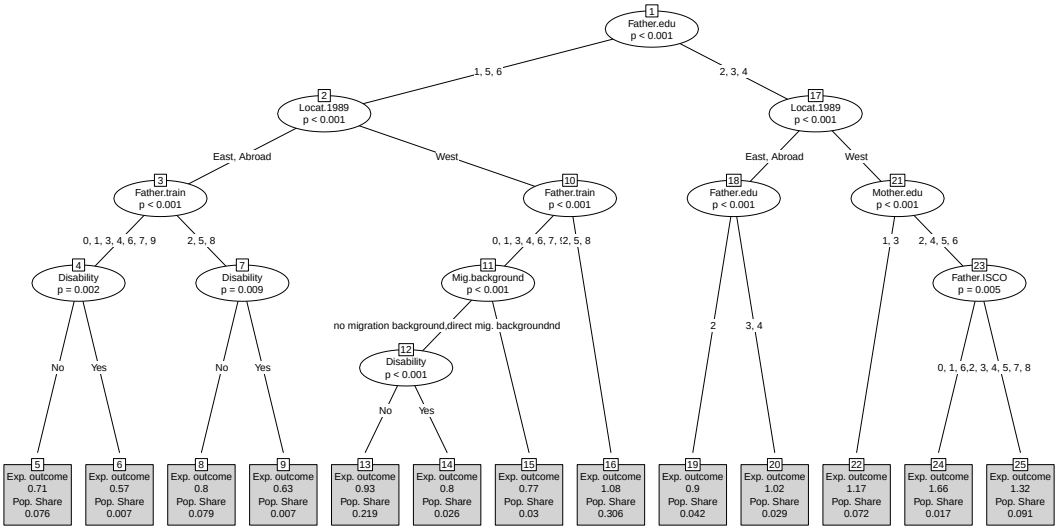


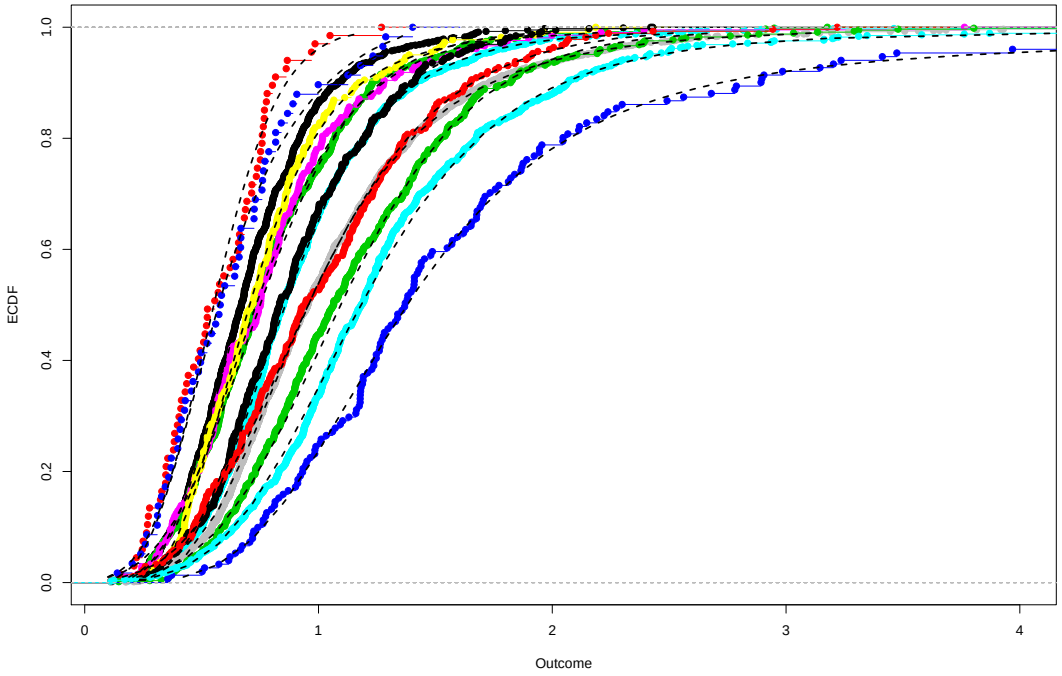
SOEPv33, 2003.



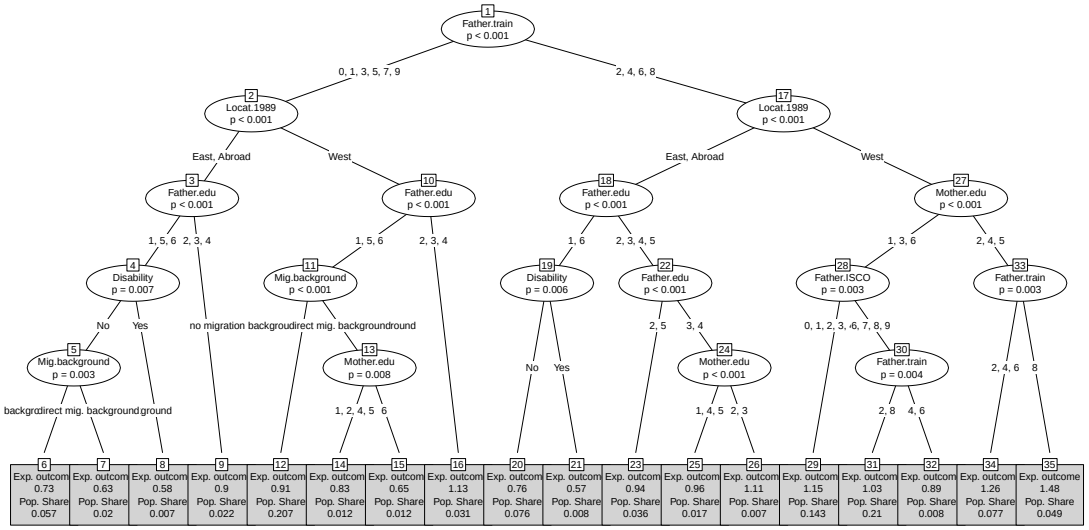


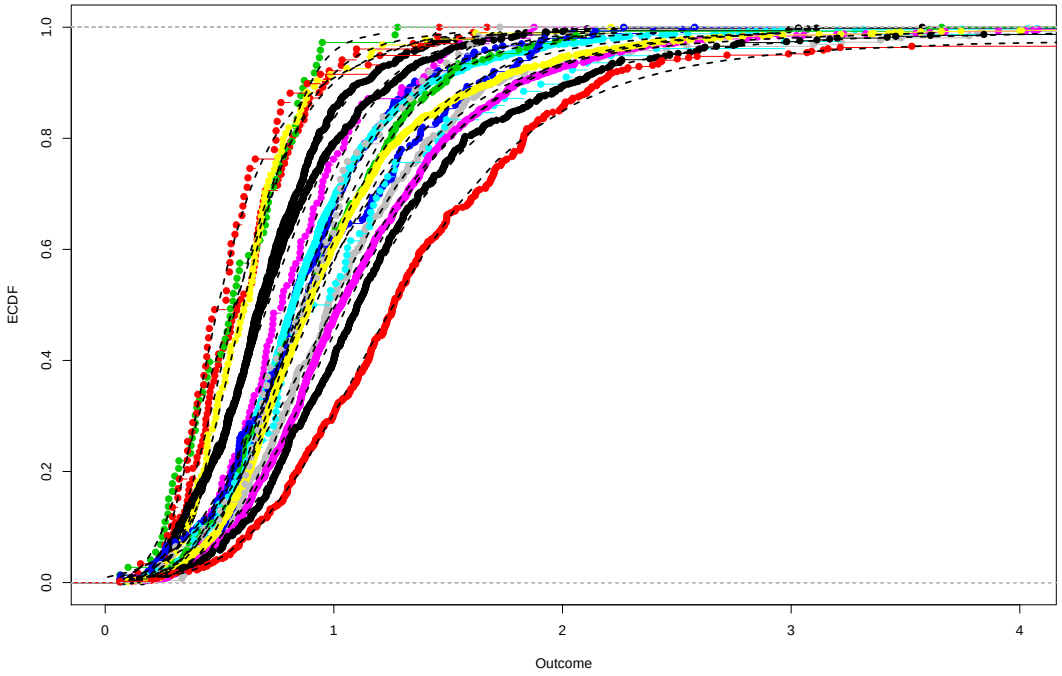
SOEPv33, 2004.



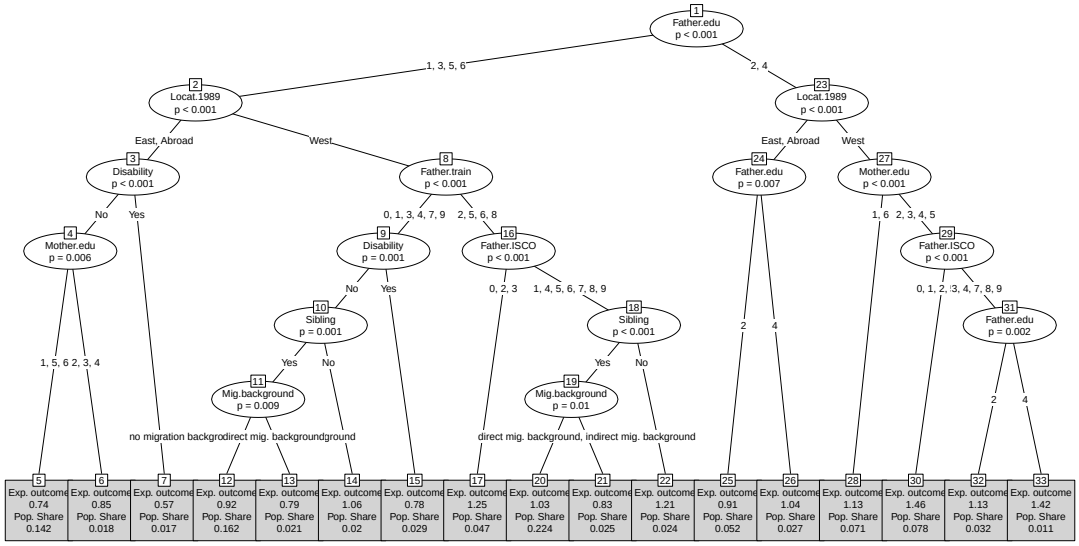


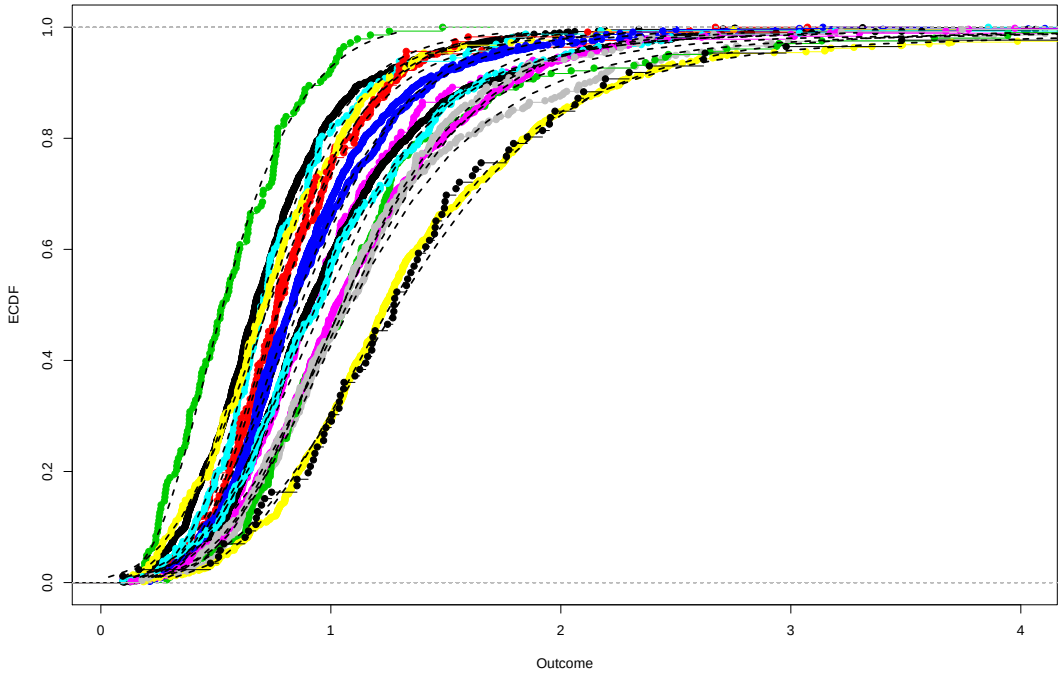
SOEPv33, 2005.



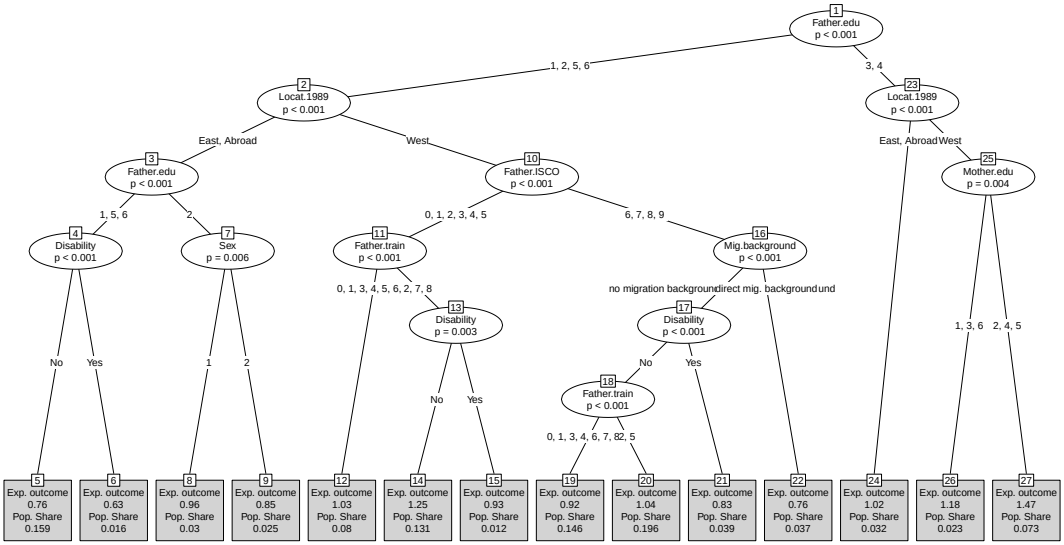


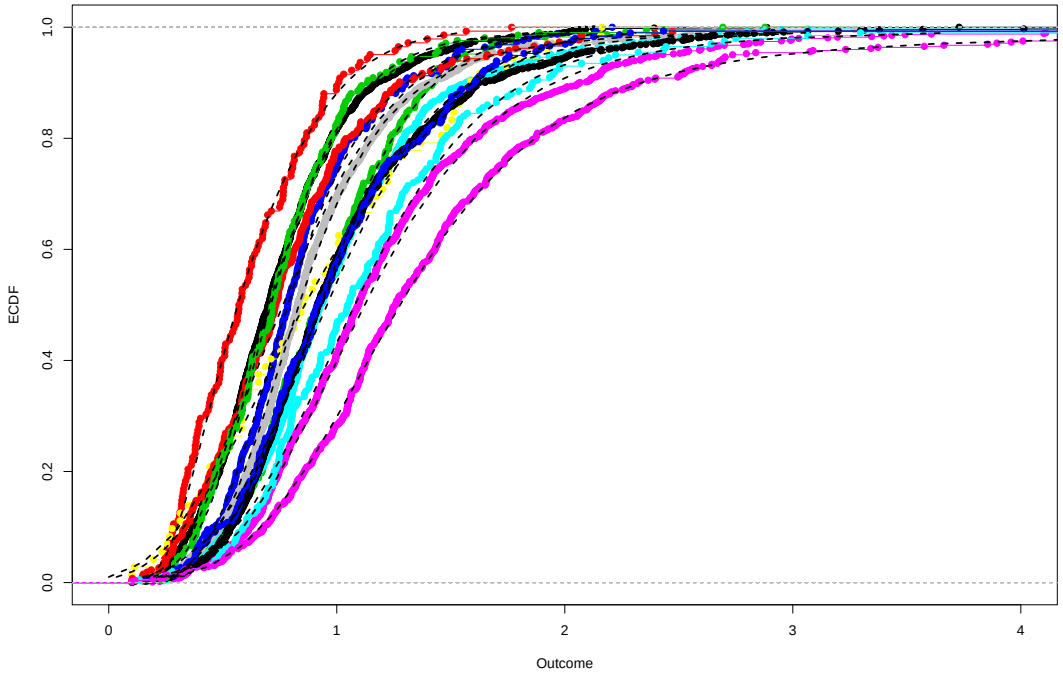
SOEPv33, 2006.



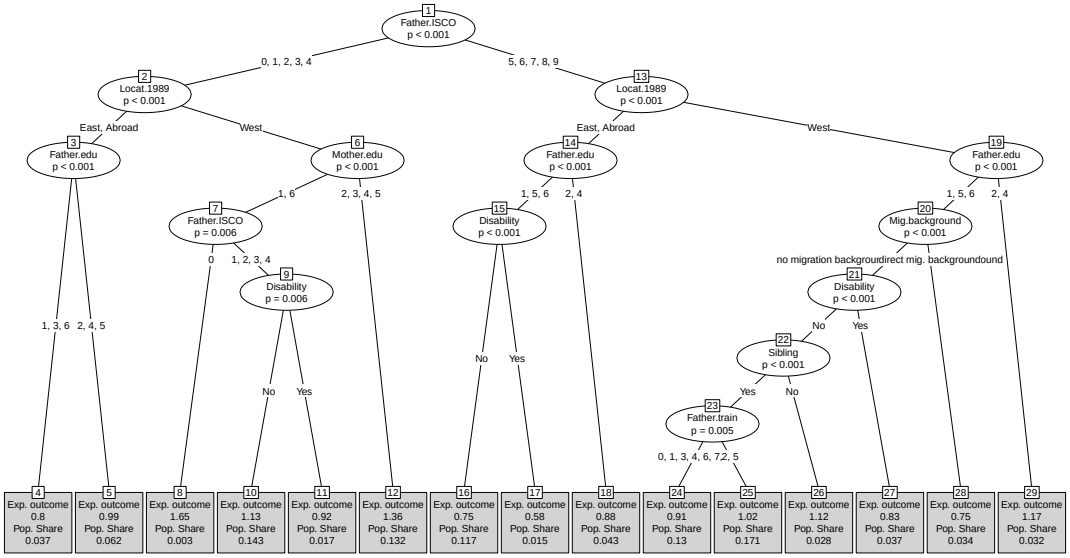


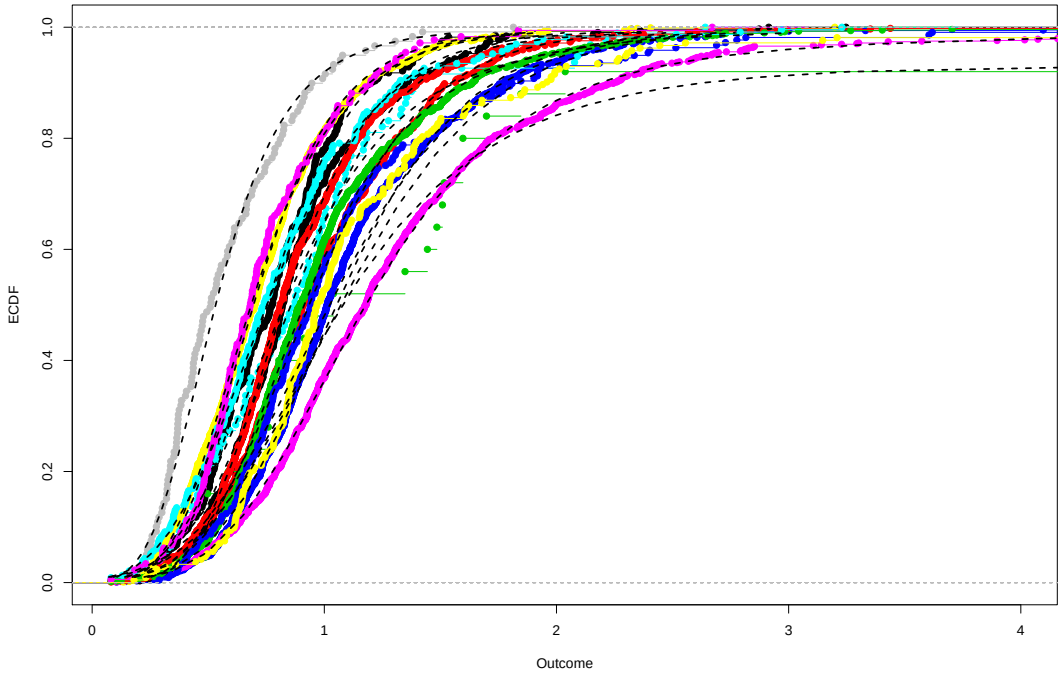
SOEPv33, 2007.



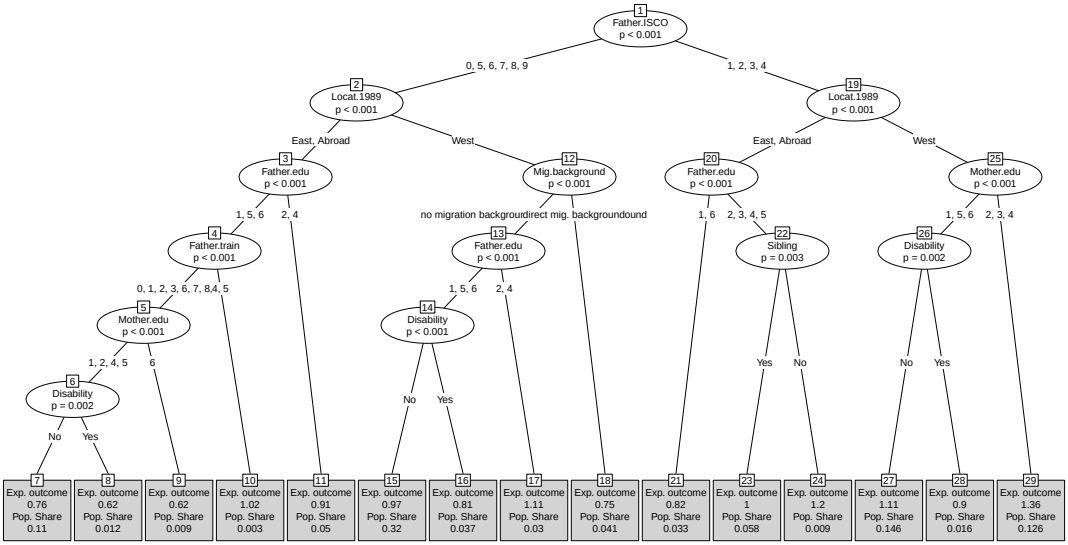


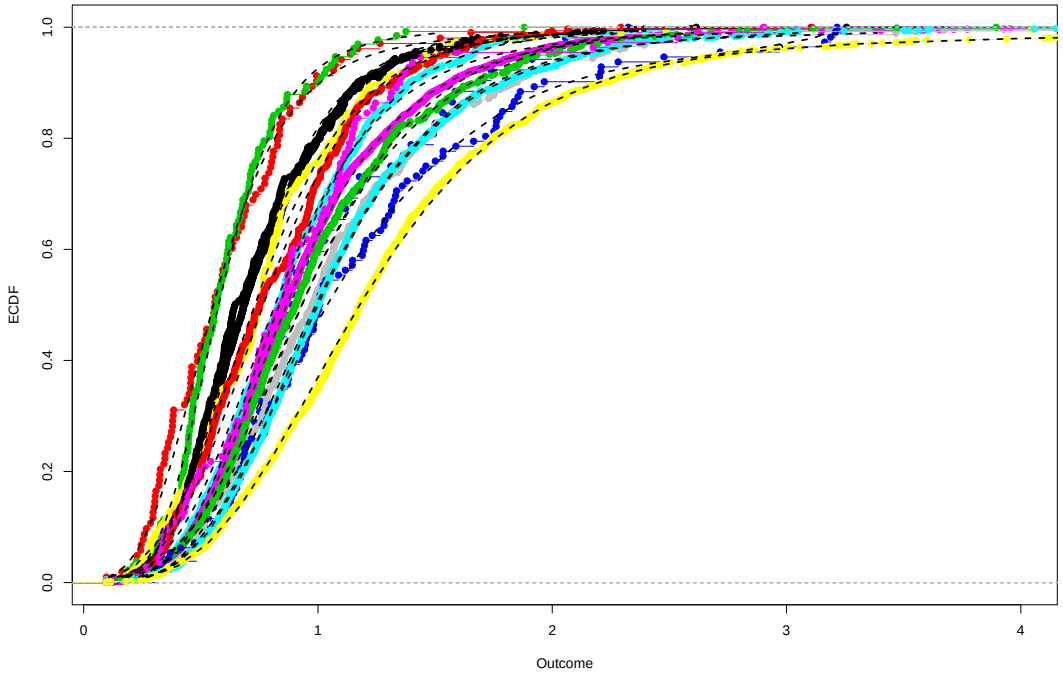
SOEPv33, 2008.



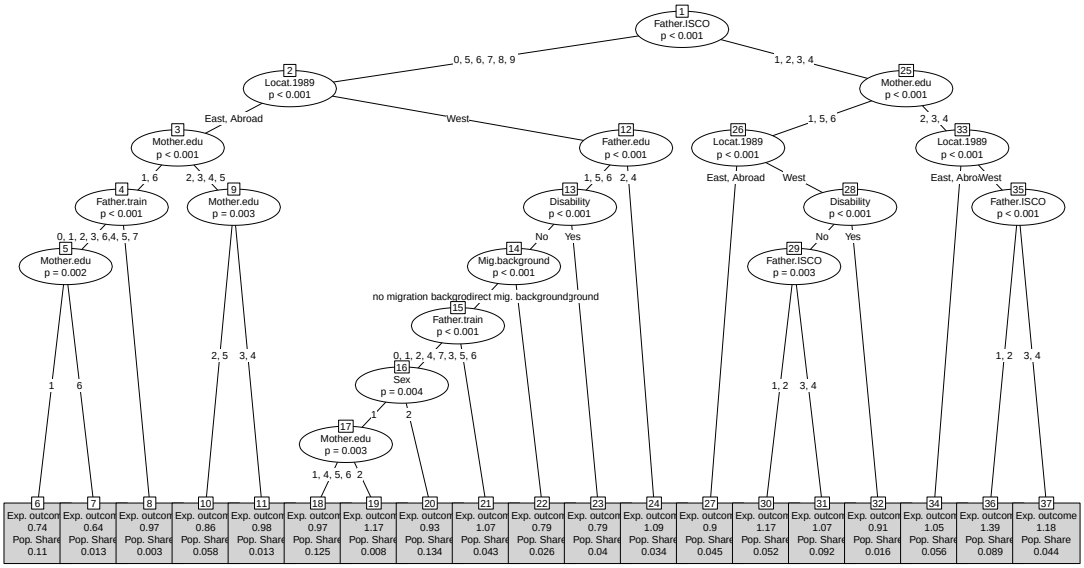


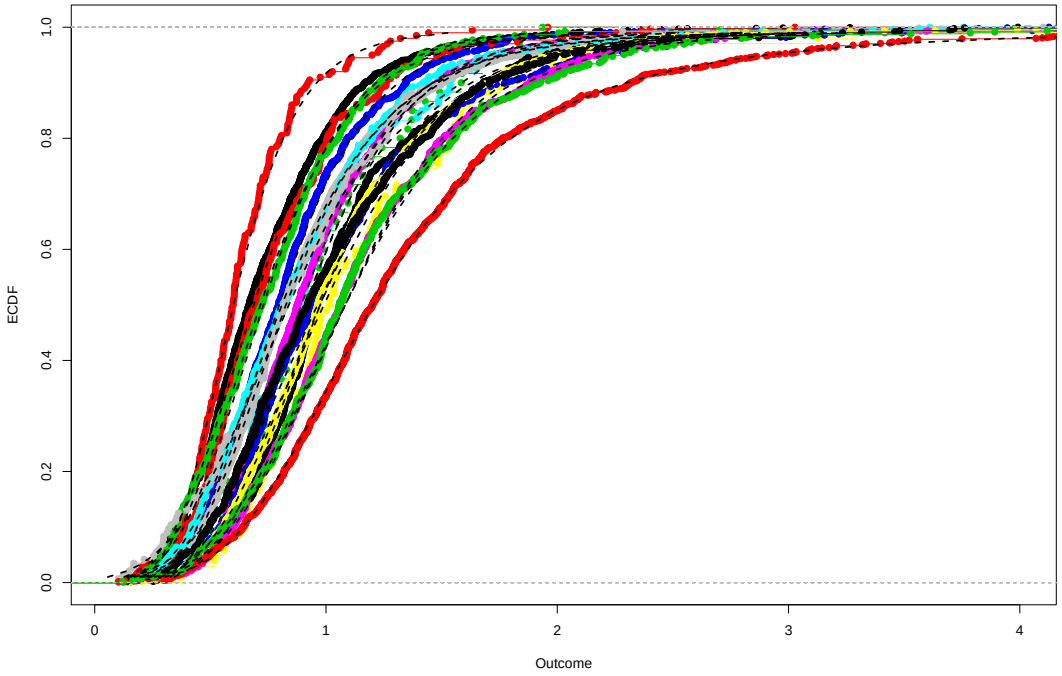
SOEPv33, 2009.



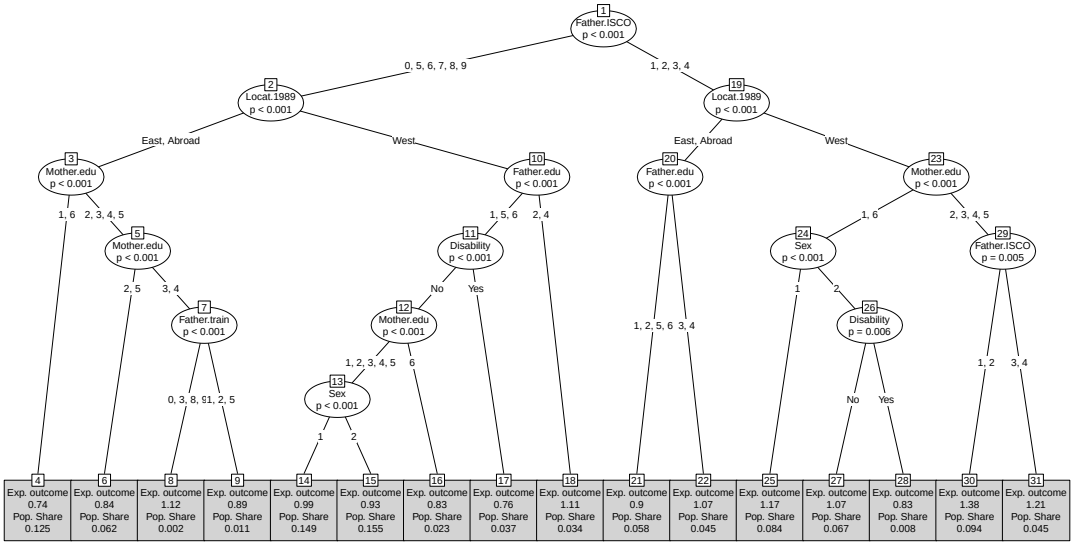


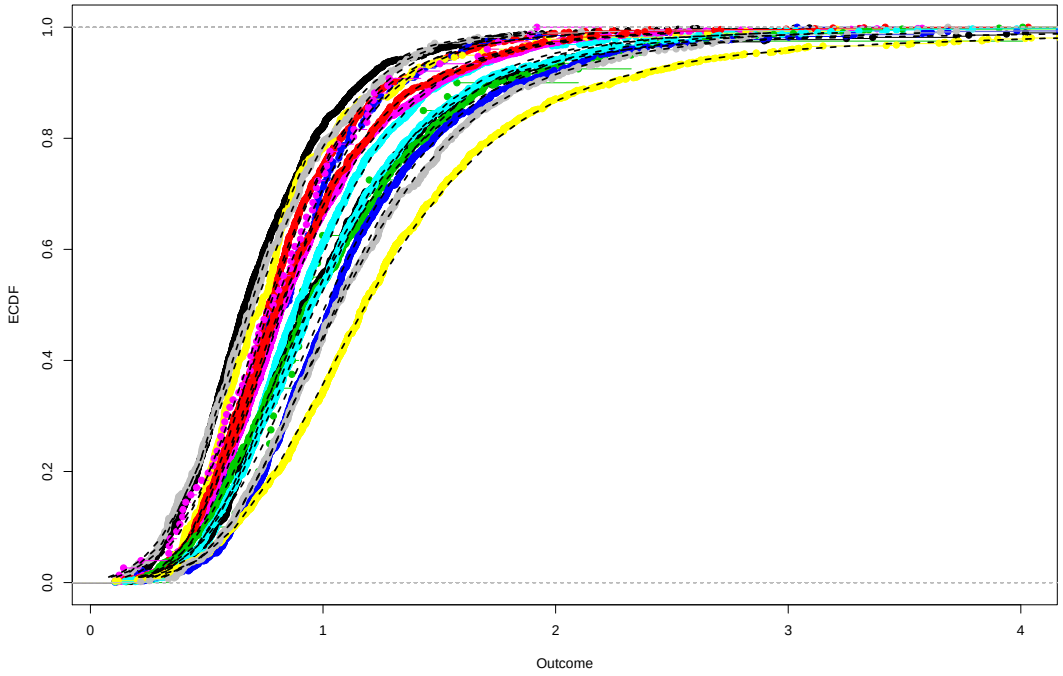
SOEPv33, 2010.



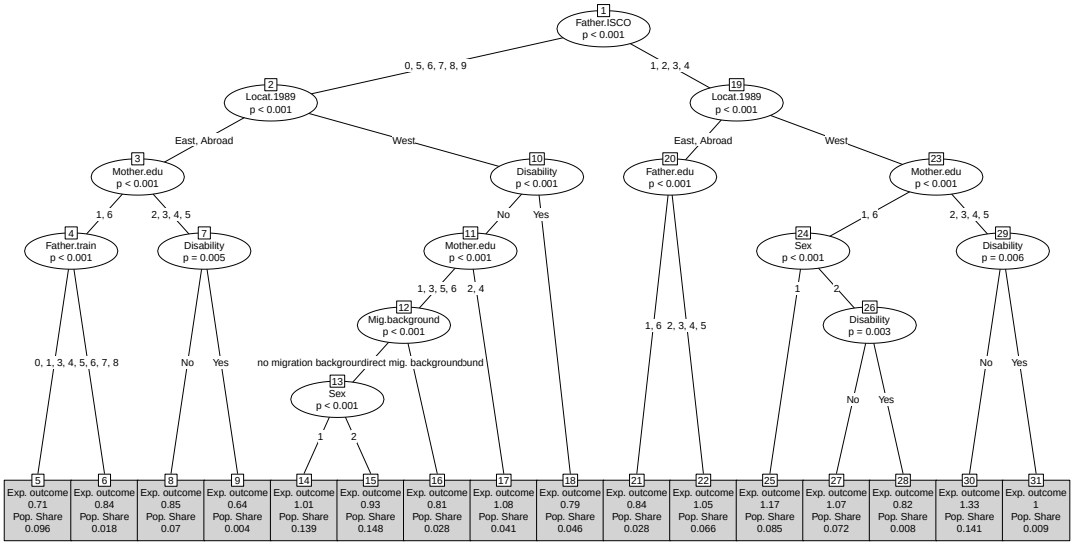


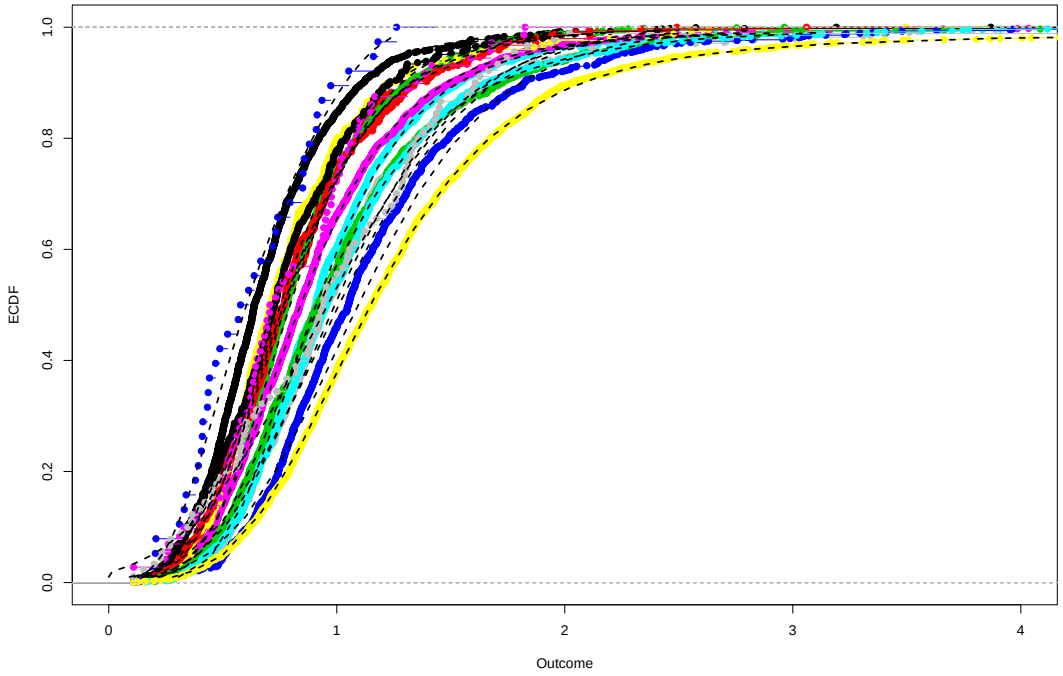
SOEPv33, 2011.



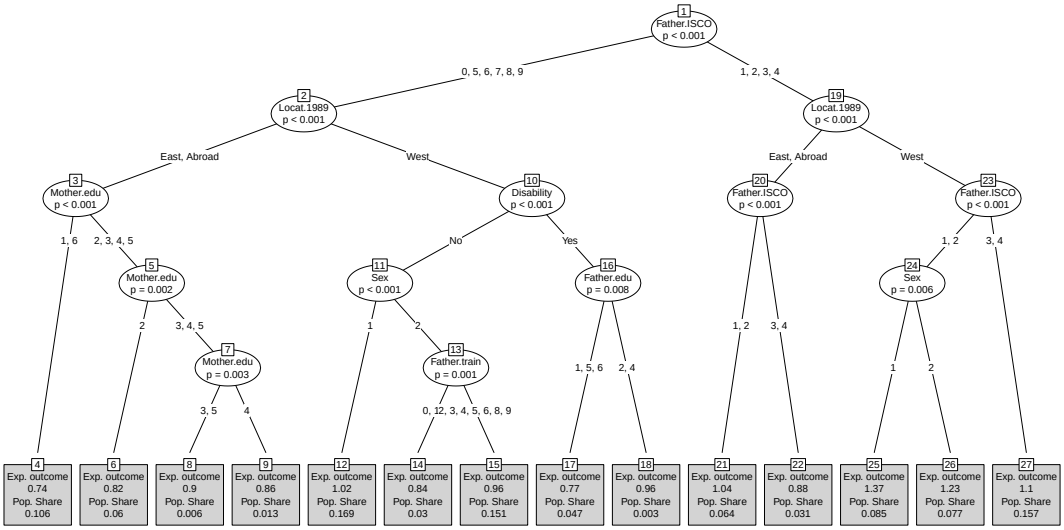


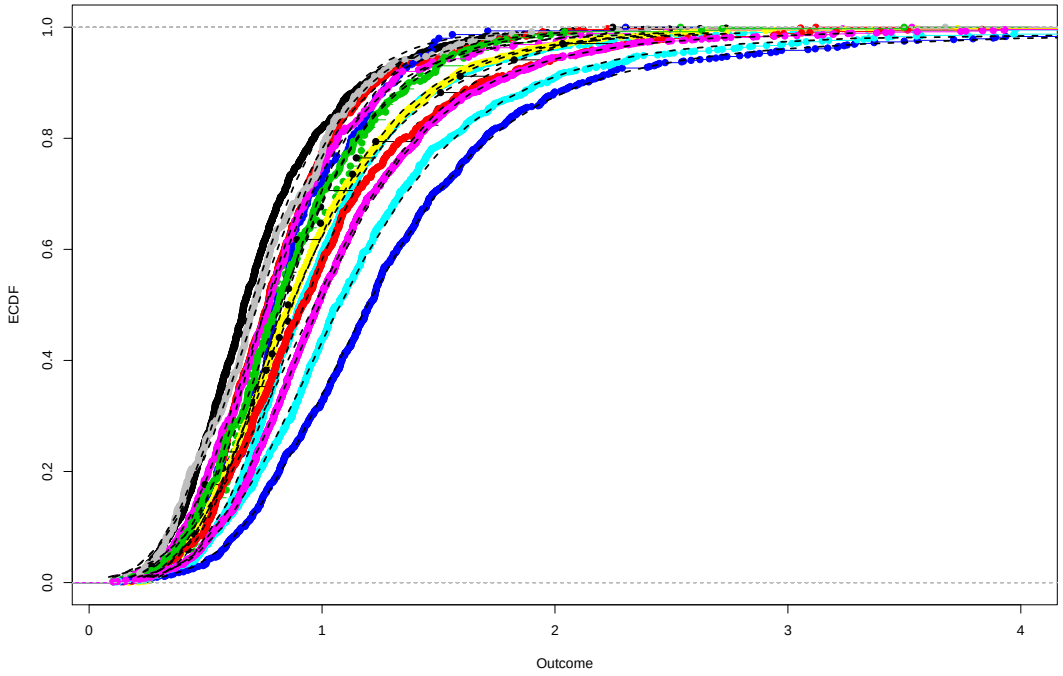
SOEPv33, 2012.



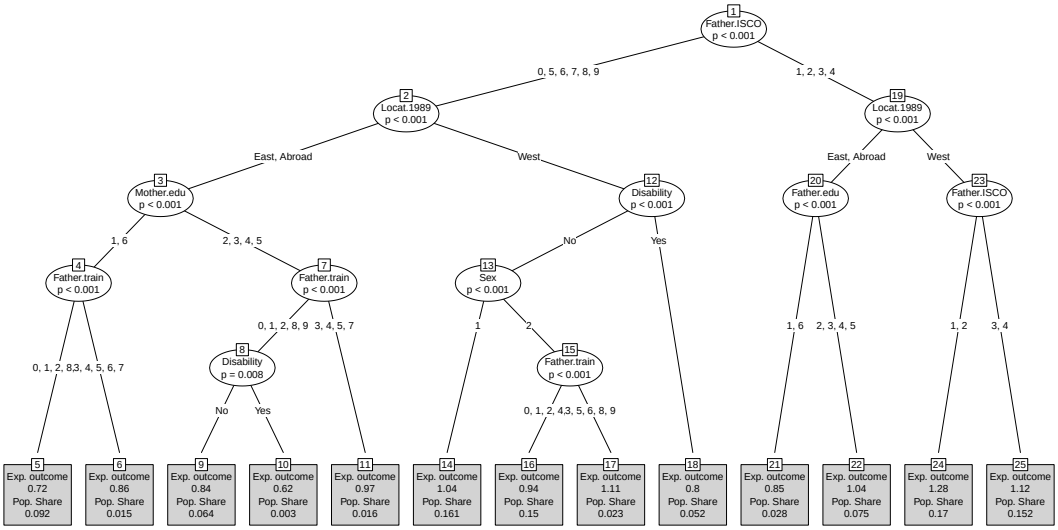


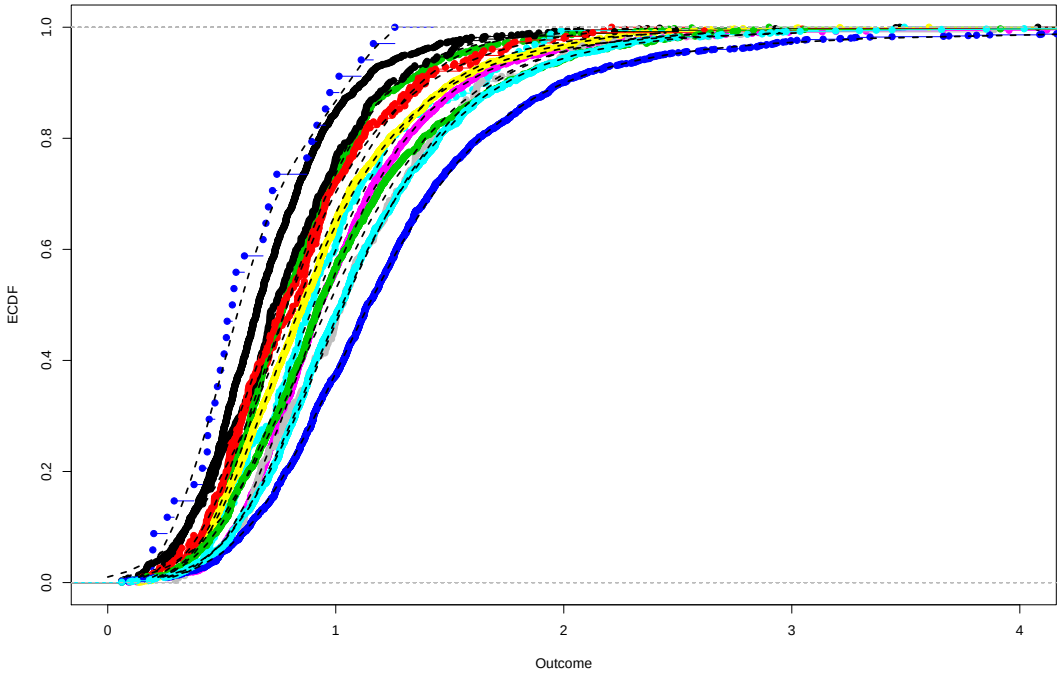
SOEPv33, 2013.





SOEPv33, 2014.





SOEPv33, 2015.