Are exposed tree roots a bad thing? Pros and cons of using machine learning to study social inequalities

JRC Centre for Advanced Studies

4th Annual Workshop 2021

Paolo Brunori

III - London School of Economics

Material

- with Paul Hufe and Danile Mahler, *The roots of inequality: estimating inequality of opportunity from regression trees and random forest* (in progress);

- with Guido Niedhöfer, *The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach* (ROIW, 2021);

- with Davillas, Jones and Scarchilli, *Model-based Recursive Partitioning to Estimate Unfair Health Inequalities in the United Kingdom* (in progress).

Motivation

- Machine Learning is increasingly used in social science;
- is ML useful to analyse socially inequalities?
- trees-based algorithms may unveil the roots of inequality;
- trees with exposed roots tend to be weak;
- but there are precautions we can take.

Equality of opportunity

- Moral philosophers: Rawls (1971), Sen (1979), Dworkin (1981), Arneson (1989), Cohen (1989);

- John Roemer:
  Equality of opportunity (1998);

- Marc Fleurbaey and Francois Maniquet:
  Responsibility-sensitive egalitarianism (2011);

- today one of the most universally accepted political ideal.

Social inequalities reproduction

Nicolas Lokhoff, Social Pyramid, 1901

Inequality of opportunity (IOP)

- ex-ante interpretation:

  IOP = between-type inequality;

- ex-post interpretation:

  IOP = inequality within individuals making same choices.

Between-type IOP

- Early contributions: Marrero and Rodriguez (2012) report IOP about 10% of total inequality in EU;

- interpreted as lower bound (Ferreira and Gignoux, 2011), inflated with longitudinal surveys (Hufe et al., 2017);

- the growing availability of data create clashes with Roemer's definition of types;

- discretionality means incomparability of estimates.

IOP measurement as a prediction problem

- A trivial descriptive exercize;
- but an exciting out-of-sample prediction challenge;
- to what extent circumstances beyond individual control are predictive of outcomes later in life?

Exposed roots *pros*

Picture: mihochannel.extend

Tree-based algorithms

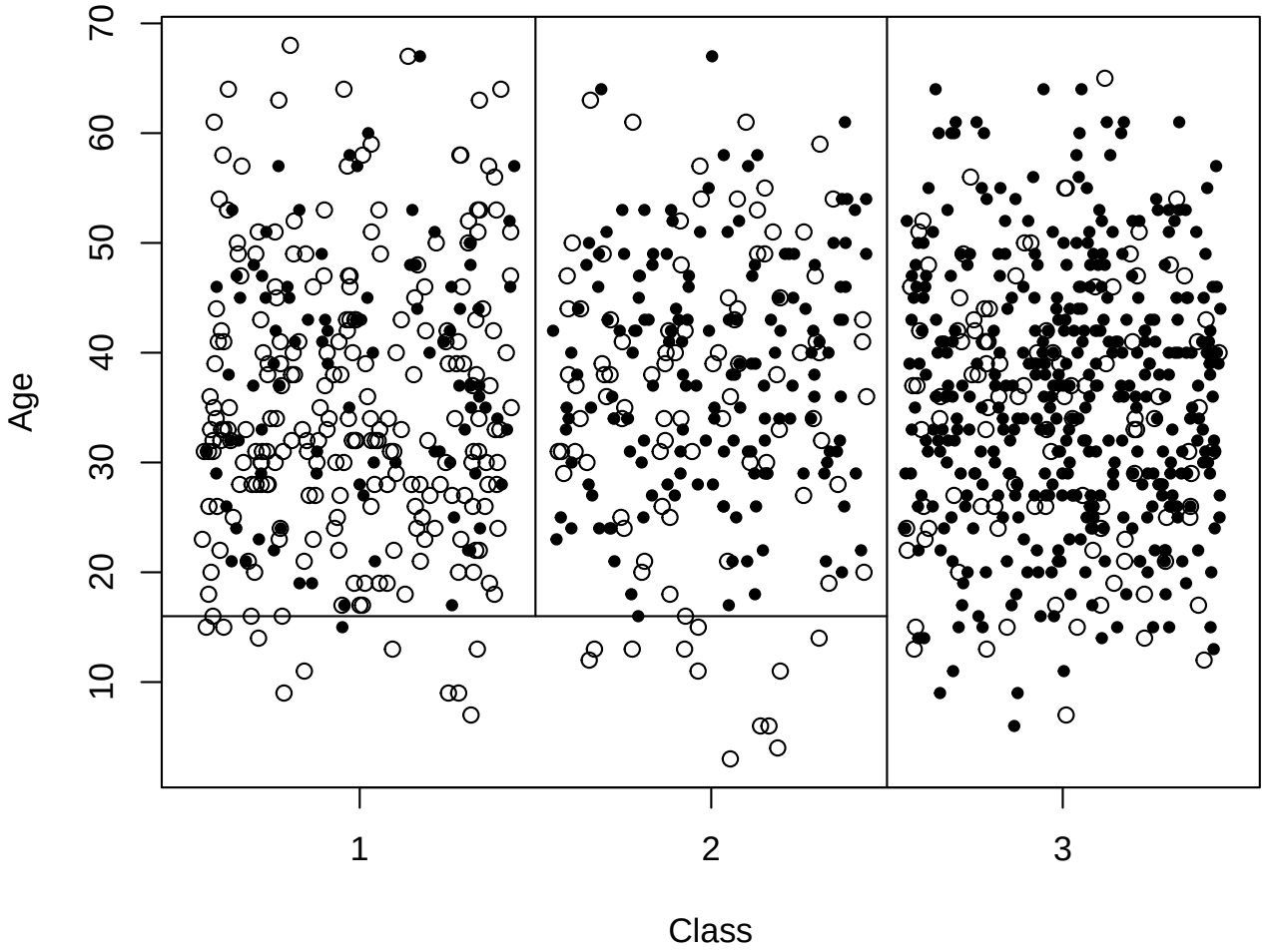
- Among supervised ML algorithms, regression trees are an attractive option to estimate between-group variability;

- trees predict a dependent variable based on observable featuress (Morgan and Sonquist,1963; Breiman et al.,1984);

- the population is divided into non-overlapping subgroups based on a partition of the predictors' space;

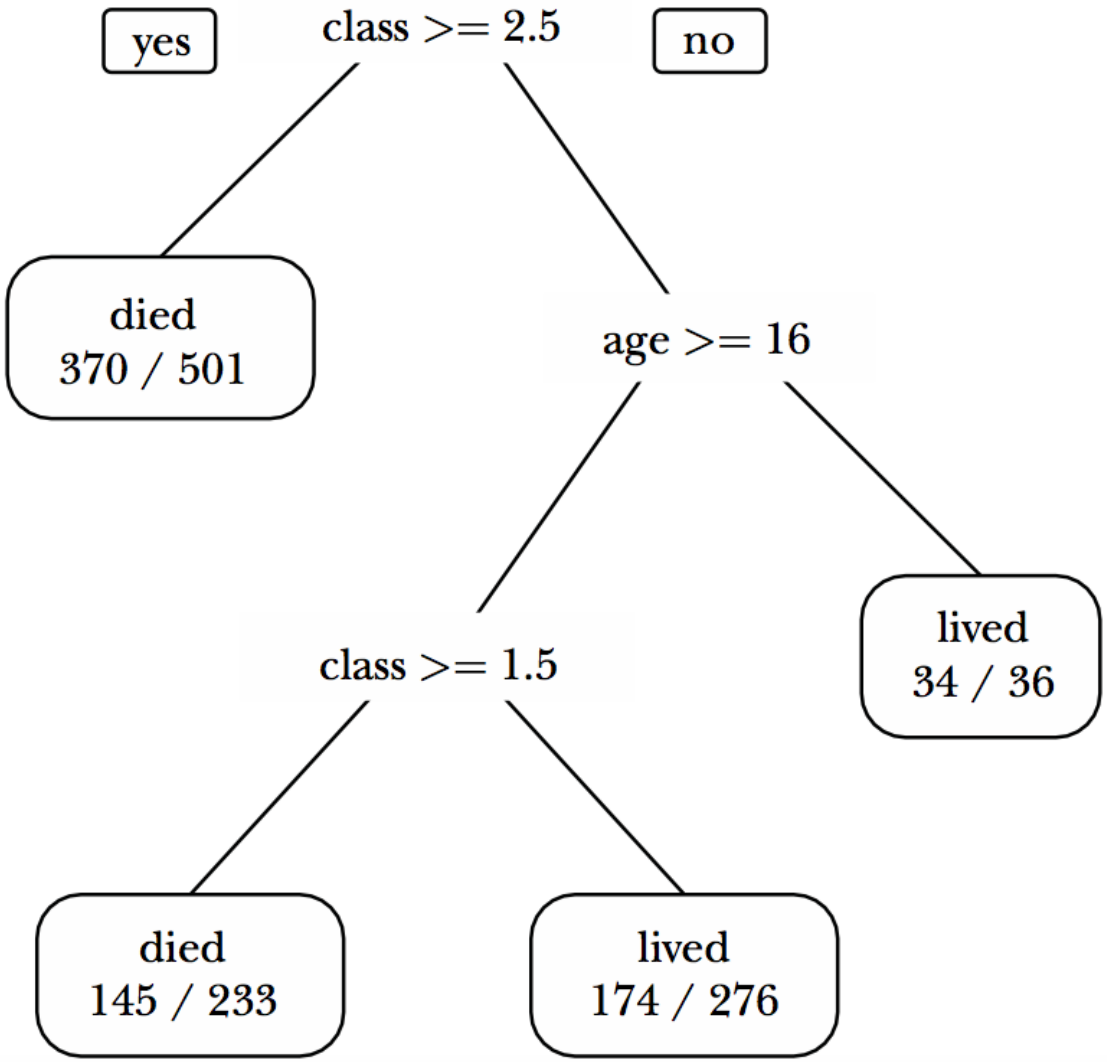
- prediction of each observation is the the mean value of the dependent variable in the group.

What is a tree? cnt.



Source: adapted from Varian, 2014

What is a tree? cnt.



Source: Varian, 2014

Pruning

- a very deep tree performs poorly out-of-sample;
- different solutions to prevent overfitting lead to different type of trees;
- *conditional inference trees* condition each split on a sequence of statistical test (Hothorn et al., 2006).

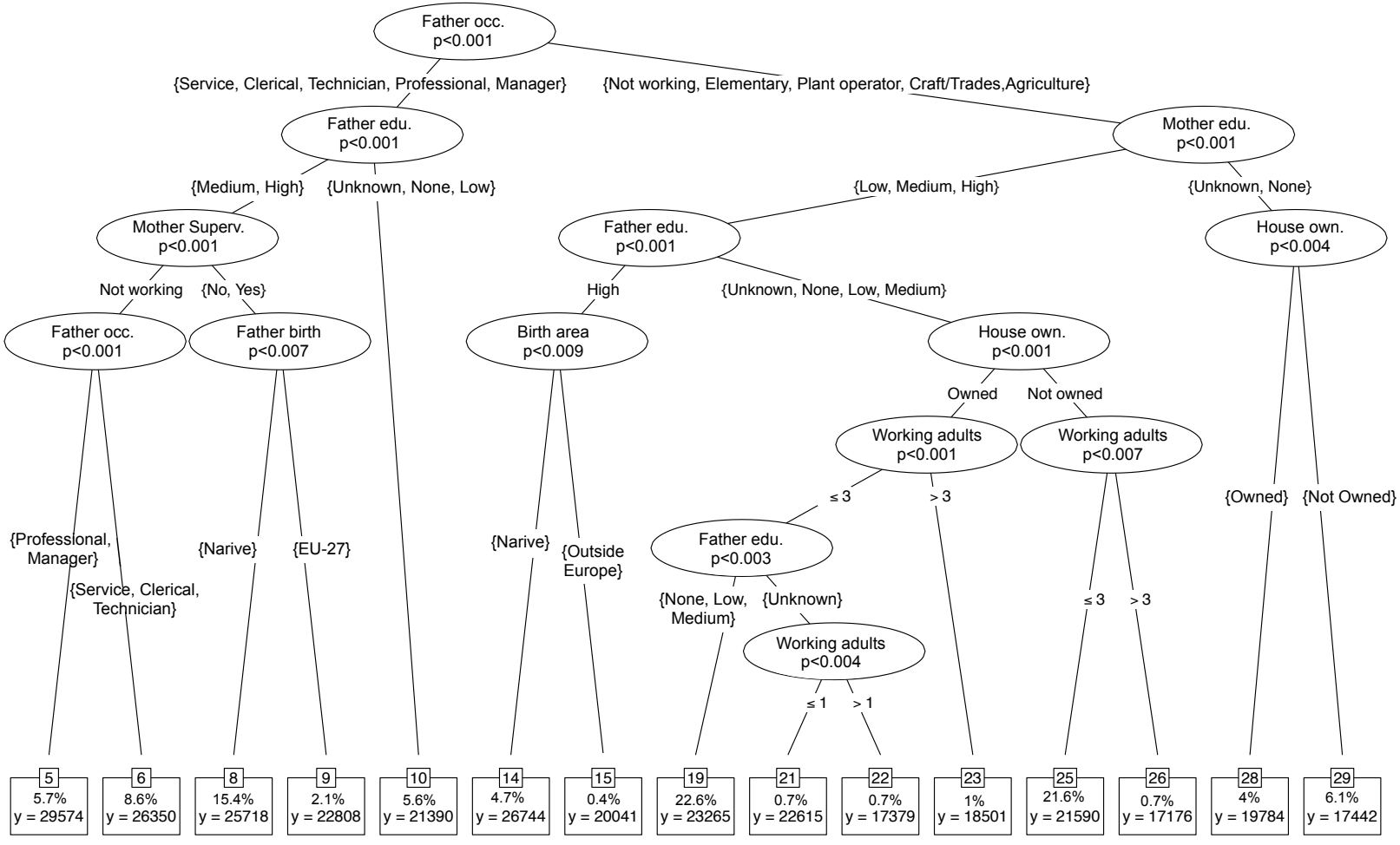
Conditional inference trees

- choose a $\alpha \in [0, 1]$;
- test the null hypothesis: outcome $\perp\!\!\!\perp$ circumstances;
- if no (adjusted) $p - value < \alpha \rightarrow$ exit the algorithm;
- select the variable, with the lowest $p - value$;
- test the discrepancy between the subsamples for each possible binary partition based on the selected variable;
- choose the splitting point that yields the lowest $p - value$;
- repeat the algorithm for each of the resulting subsamples.

Opportunity trees: *pros*

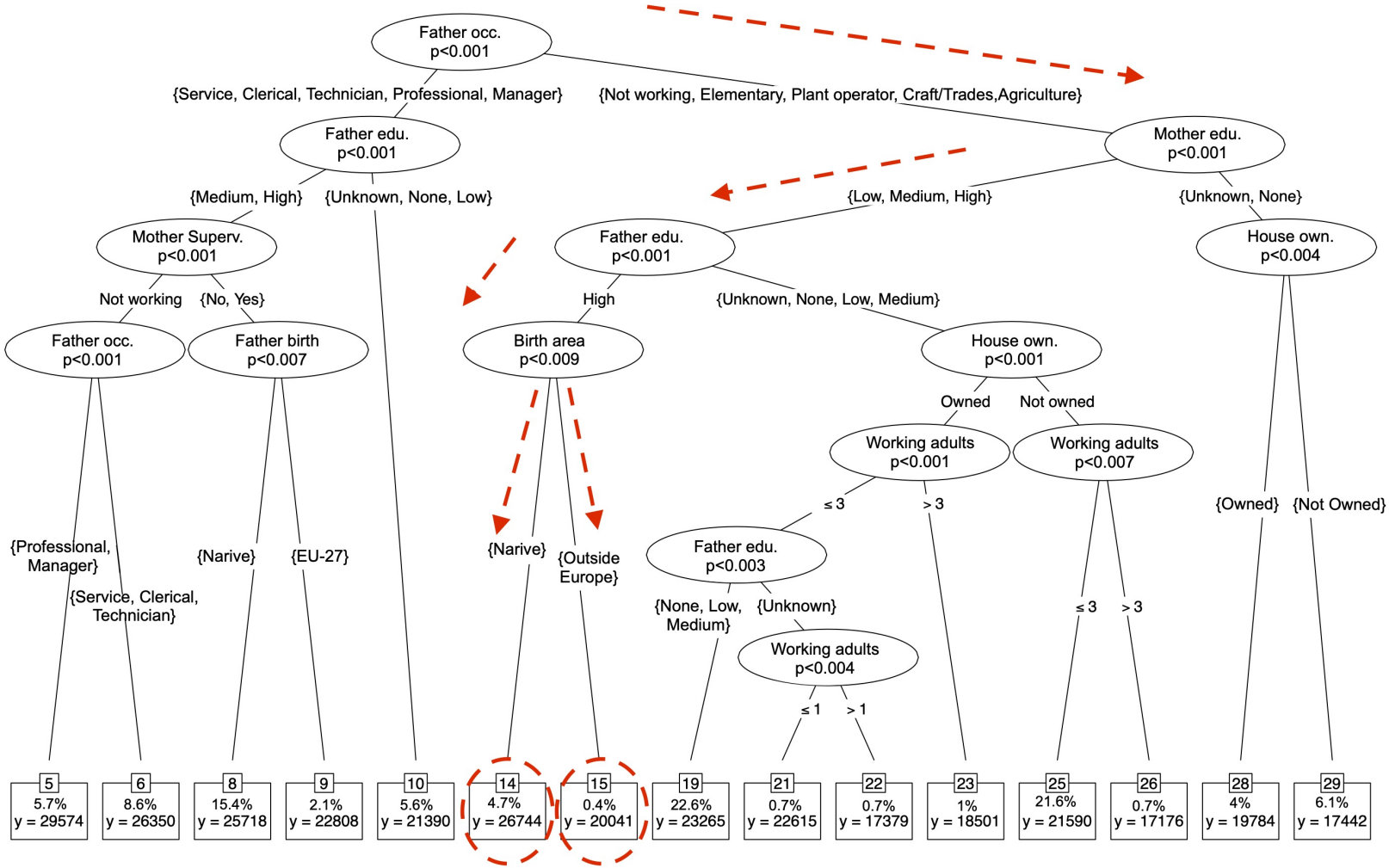
- the selection of circumstances is non-arbitrary;
- the model specification is endogenous to data;
- tell a story about the opportunity structure.

Germany, EU-SILC 2011



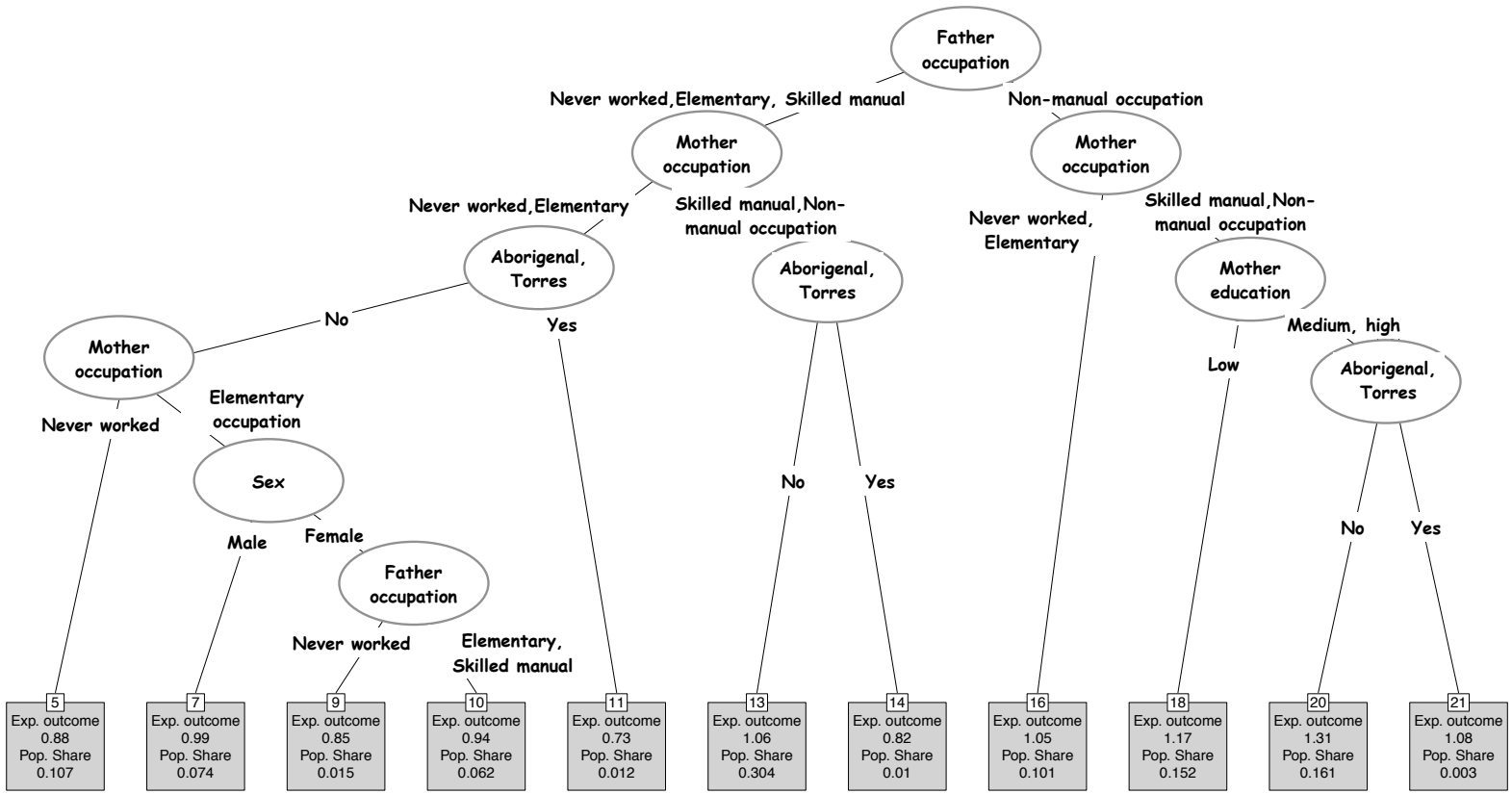
Source: Brunori, Hufe, Mahler (2020)

Germany, EU-SILC 2011

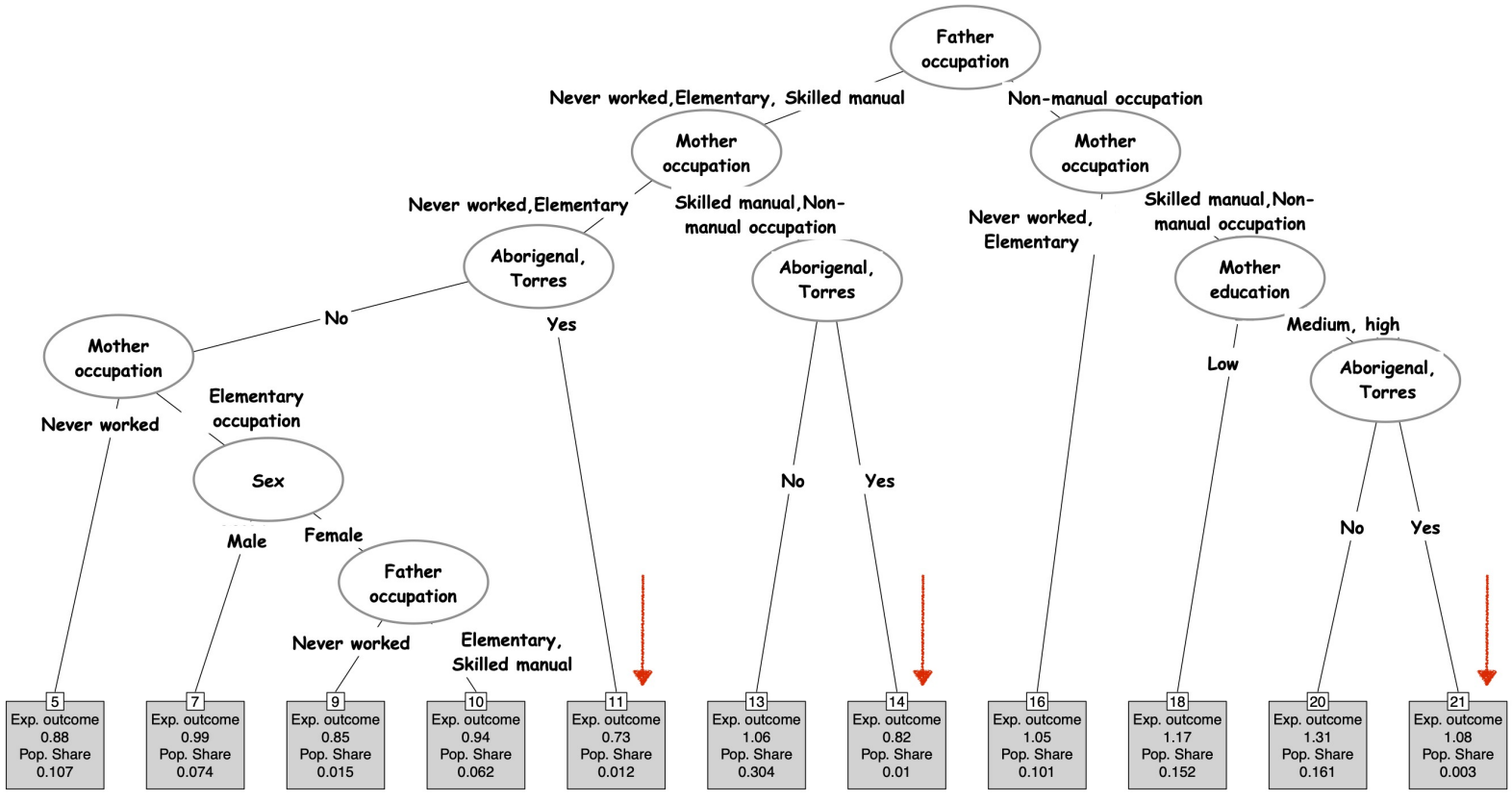


Source: Brunori, Hufe, Mahler (2020)

Australia, HILDA 2015



Australia



Source: HILDA, 2015

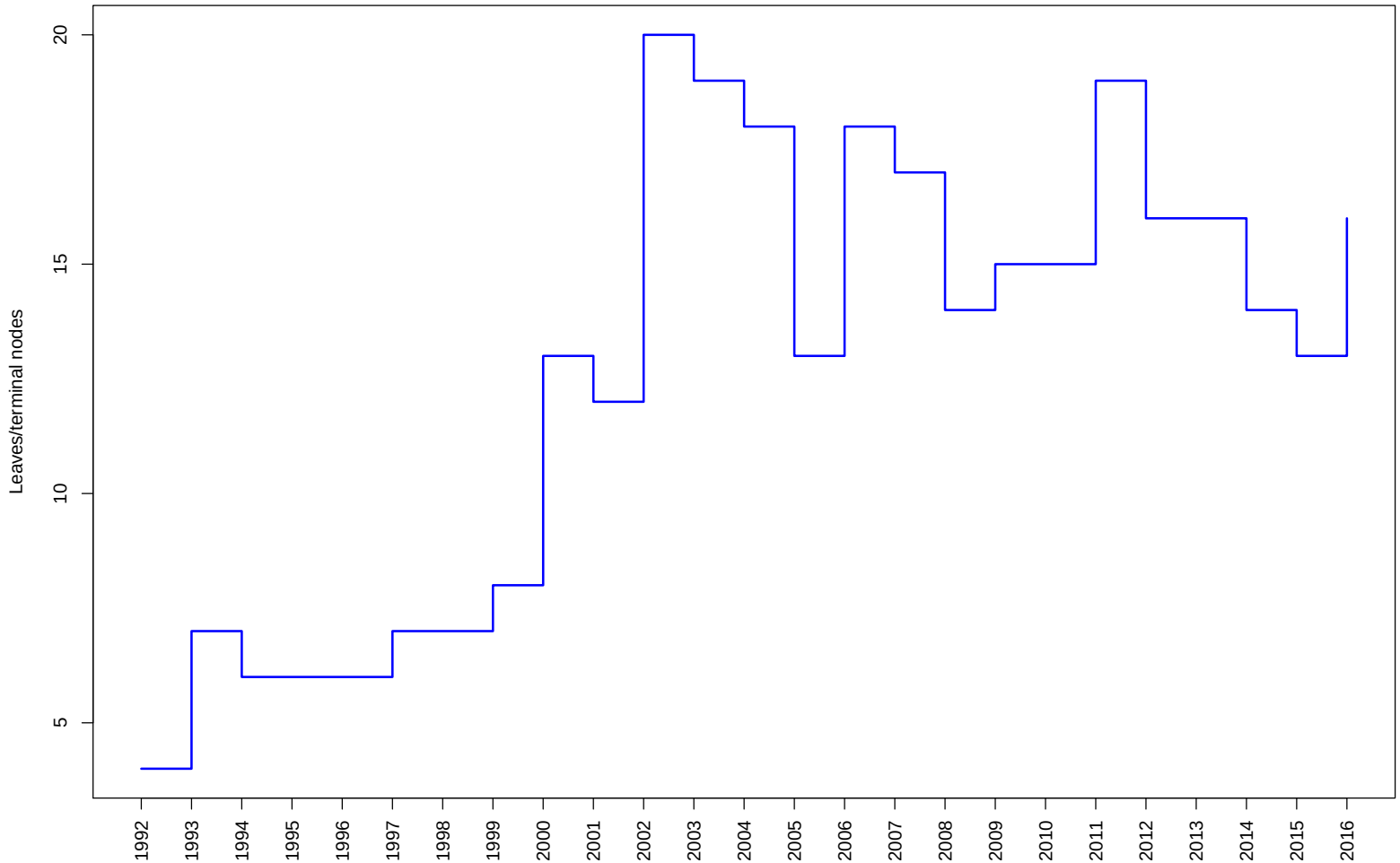
ROOT CARE: WHEN ROOTS BECOME A PROBLEM

Source: LoveYourLandscape.org

Opportunity trees: *cons*

- can be unstable (high variance);
- misleading with highly correlated circumstances (multicollinearity);
- perform poorly if the data generating process is linear.

Number of types, Germany 1992-2016

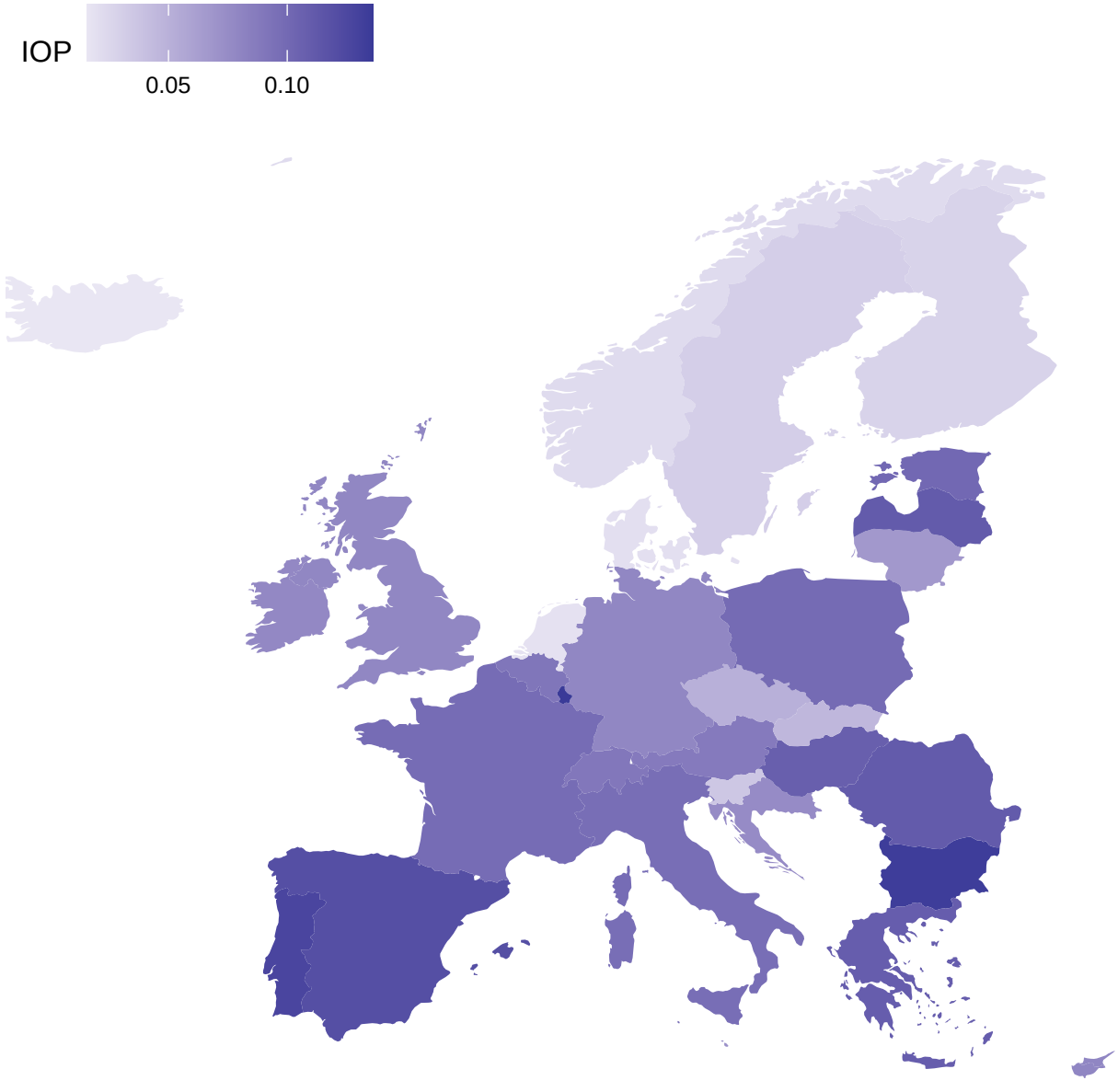


source: Brunori and Neidhöfer (2021) based on SOEP

Bootstrap aggregation and random forests

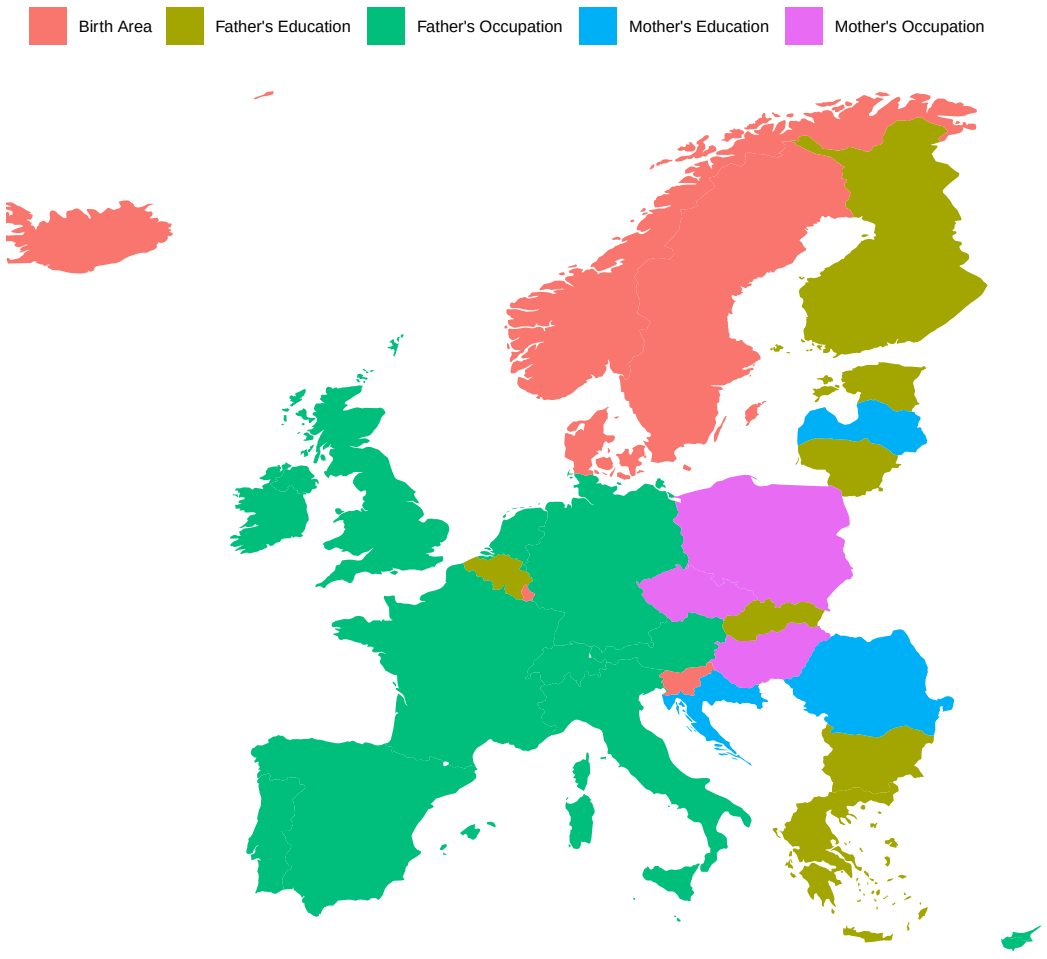
- Aggregation of many weak learning algorithms results in a stronger learner;
- a forest outperform trees by aggregating predictions of hundreds of trees;
- individual prediction is obtained from many likely opportunity structures;
- useful to assess relative predictive power of different circumstances.

Random forest: IOP in EU



source: Brunori, Hufe, Mahler, 2020.

Random forest: most predictive circumstance



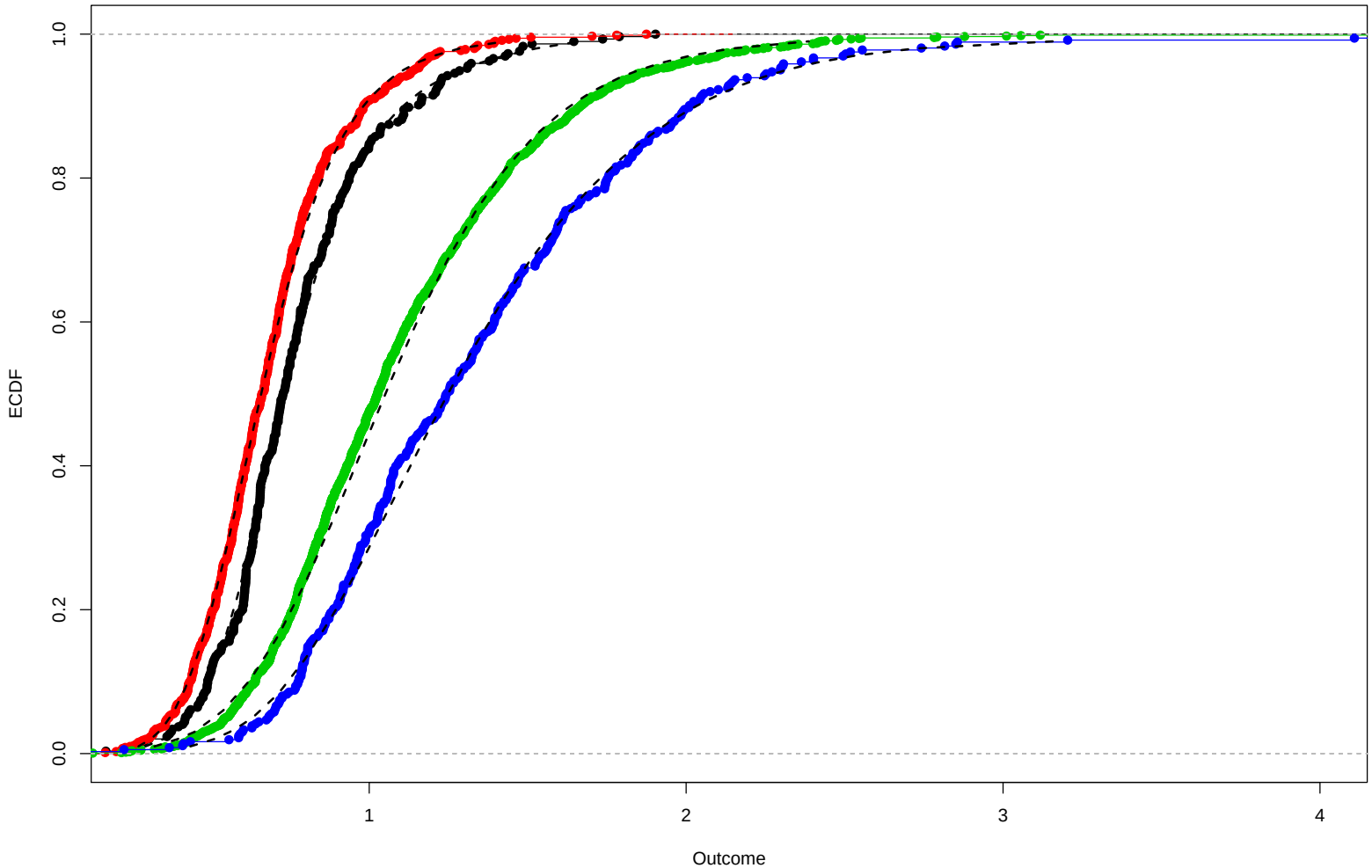
source: Brunori, Hufe, Mahler, 2020.

Did we forget choices?

- Ex-post IOP: inequality within individuals making the same choices.

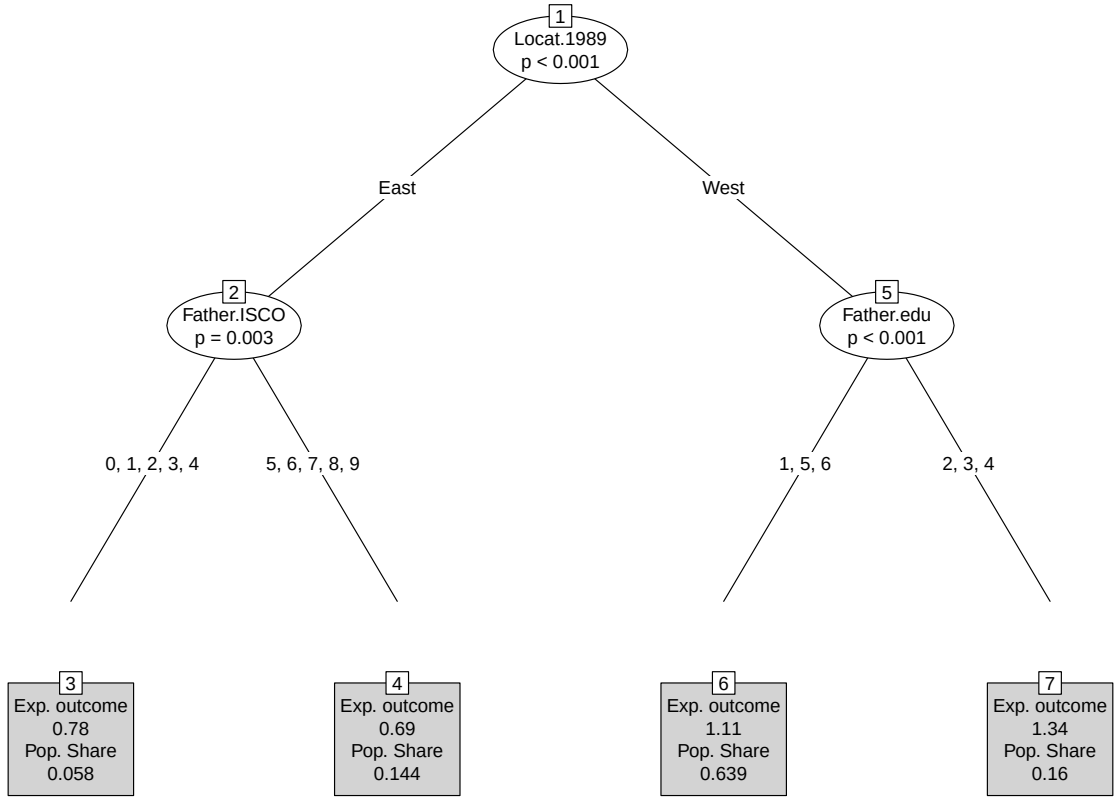
- Roemer (1998): when choices are unobservable compare type-specific outcome distributions.

Ex-post IOP: Germany 1992



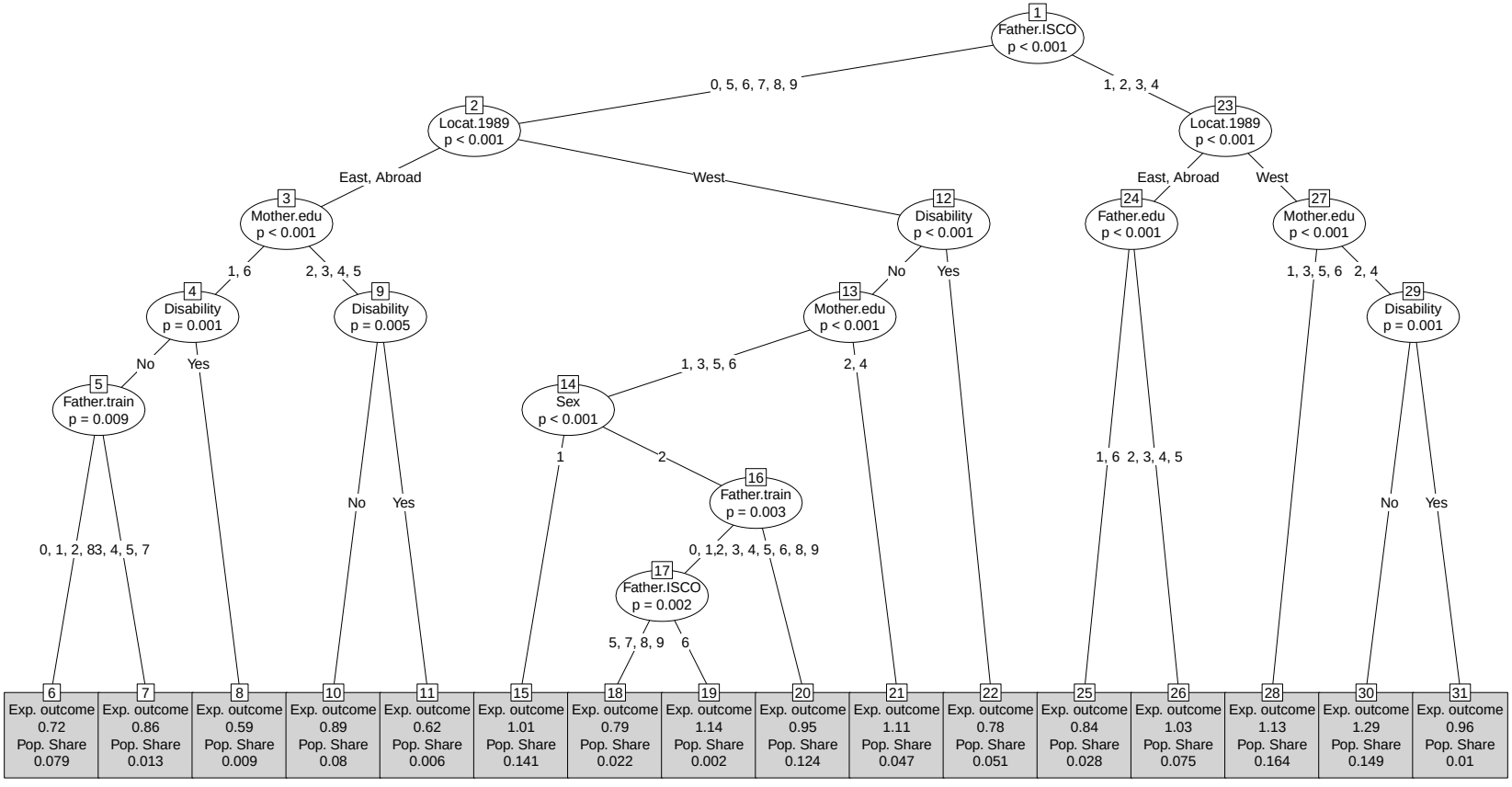
source: Brunori and Neidhöfer (2021)

Opportunity tree: Germany 1992



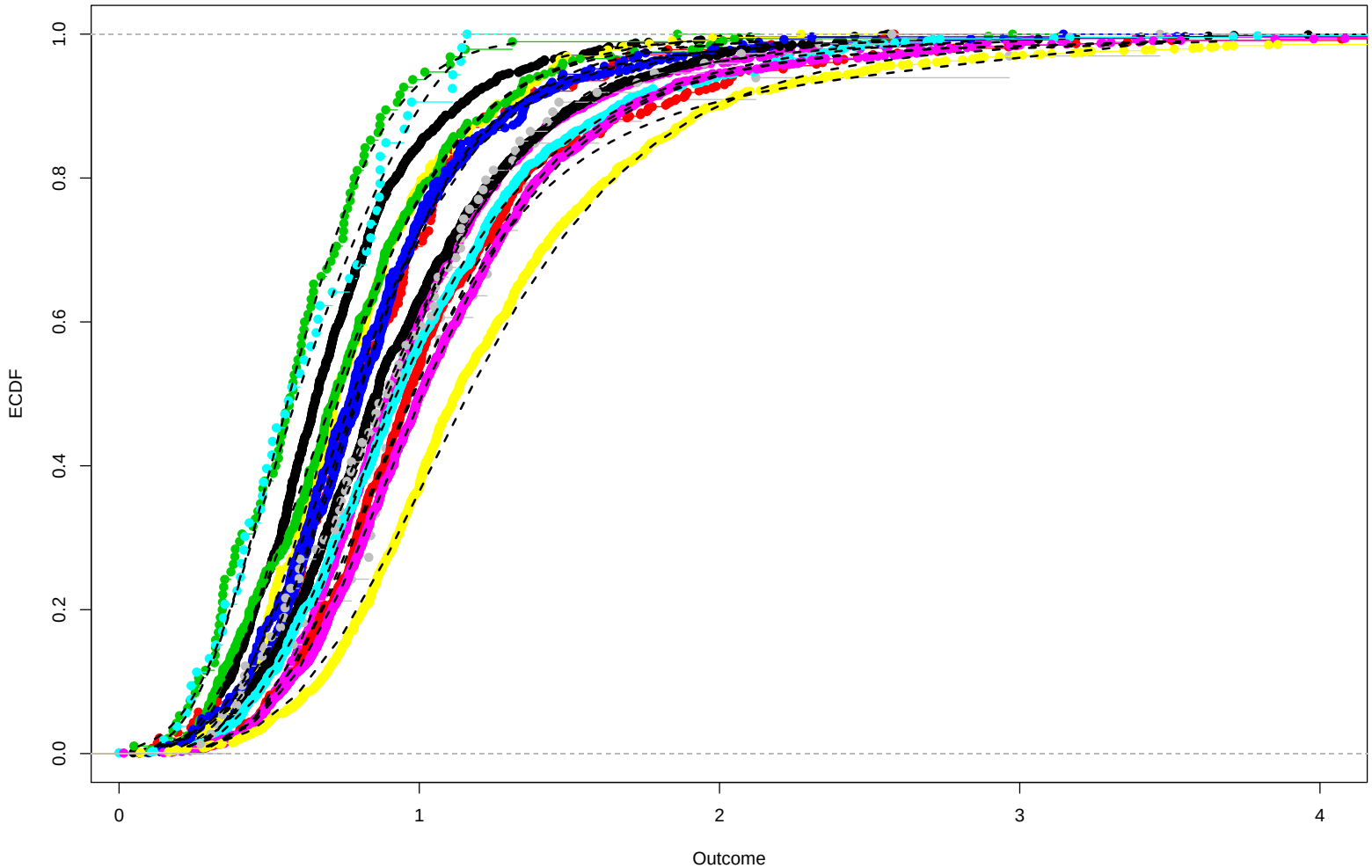
source: Brunori and Neidhöfer (2021)

Opportunity tree: Germany 2016



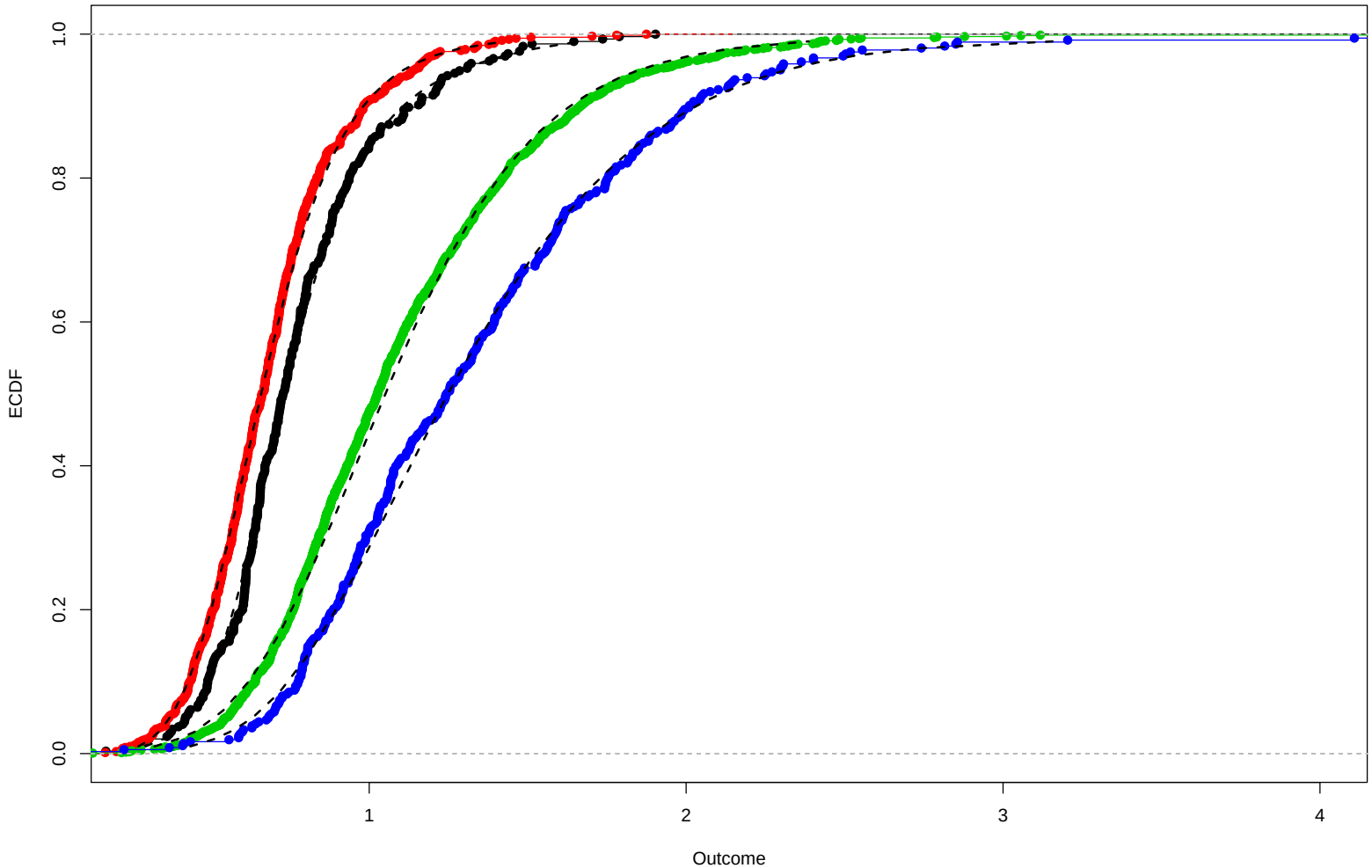
source: Brunori and Neidhöfer (2021)

Ex-post IOP: Germany 2016



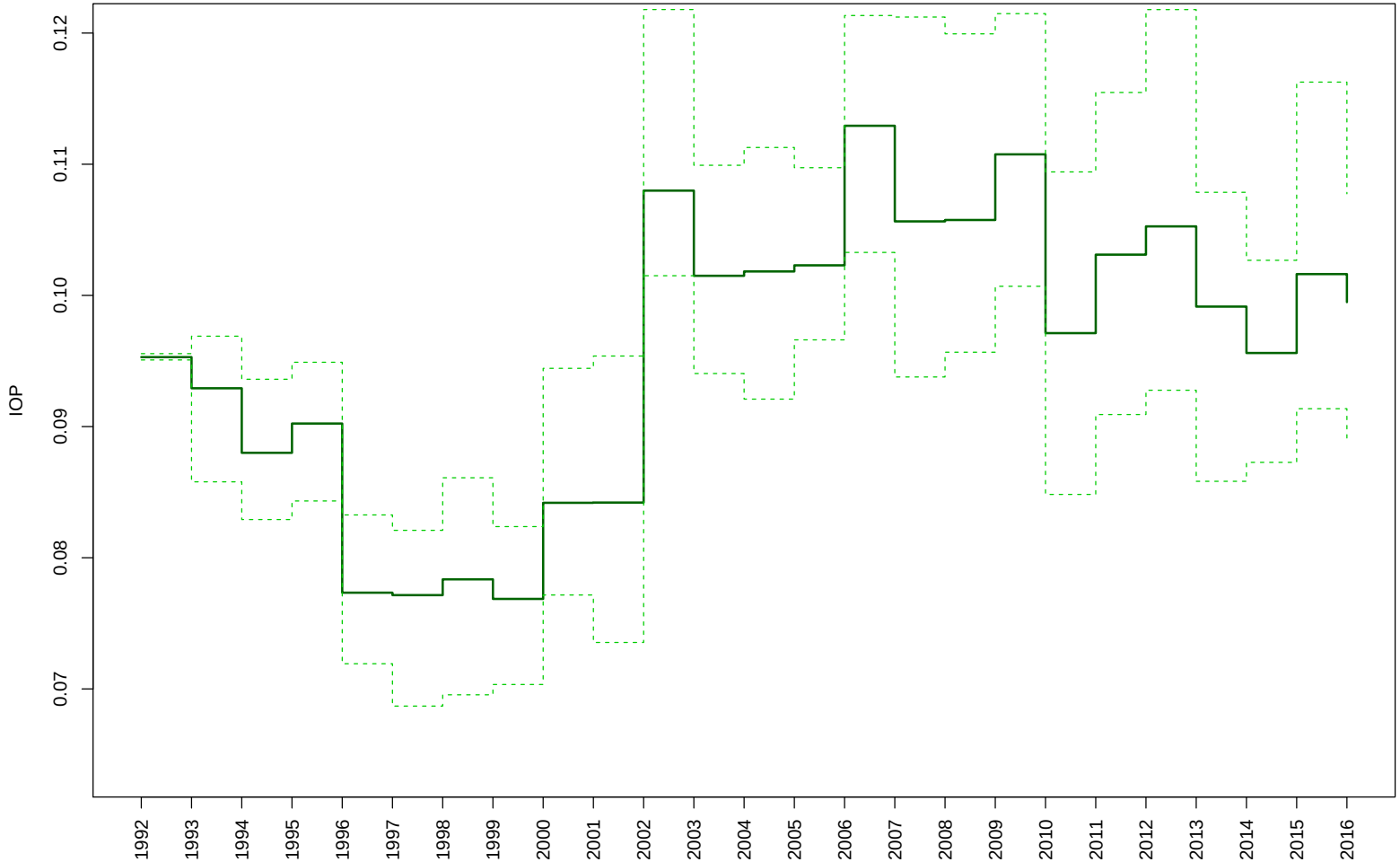
source: Brunori and Neidhöfer (2021)

Ex-post IOP: Germany 1992



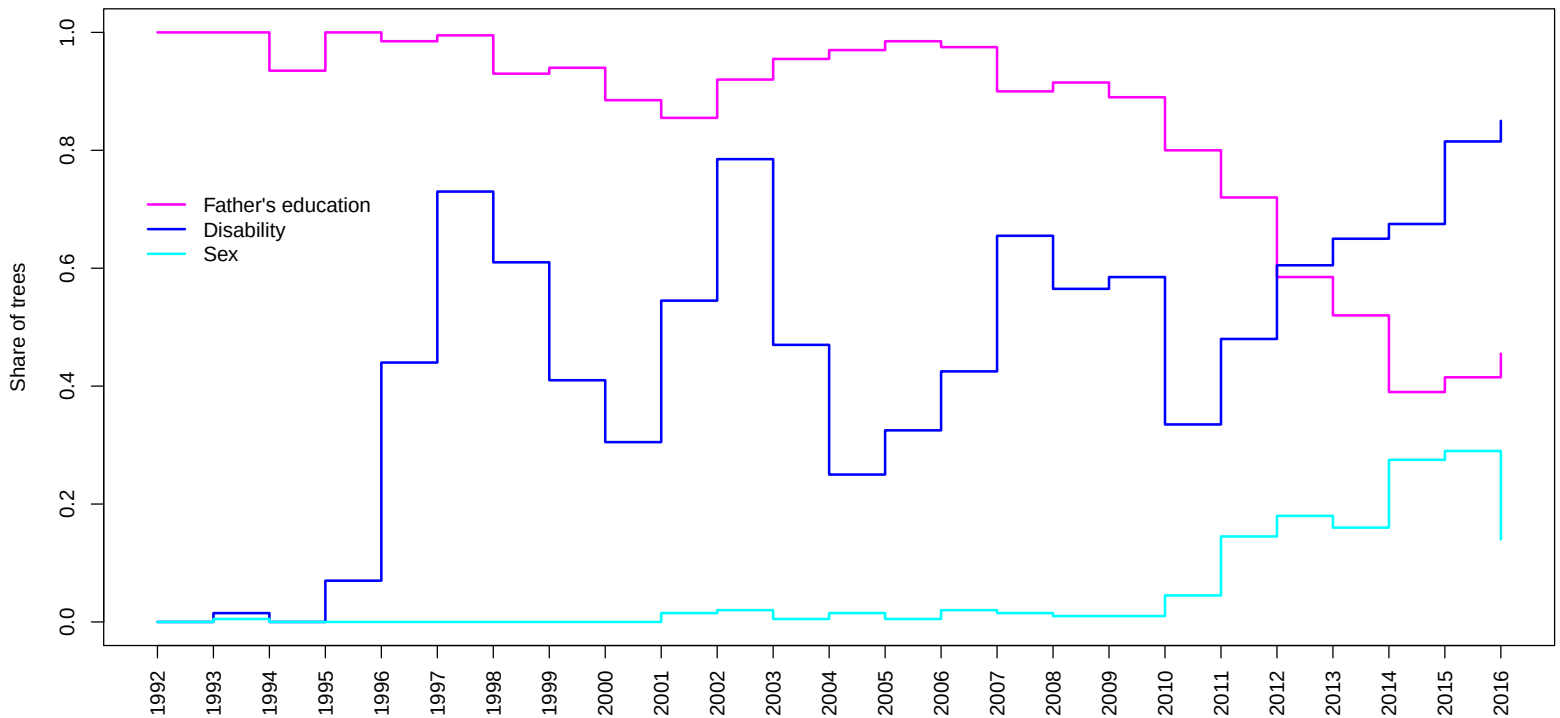
source: Brunori and Neidhöfer (2021)

IOP in Germany 1992-2016



source: Brunori and Neidhöfer (2021)

Changing opportunity structure, Germany 1992-2016



source: Brunori and Neidhöfer (2021)

Did we forget choices?

- Health inequality is a case in which observing several choices is possible;

- circumstances can affect outcome through two channels:

  1. fixed type-effect,

  2. different return to choices across types;

Model-based recursive partitioning (MOB)

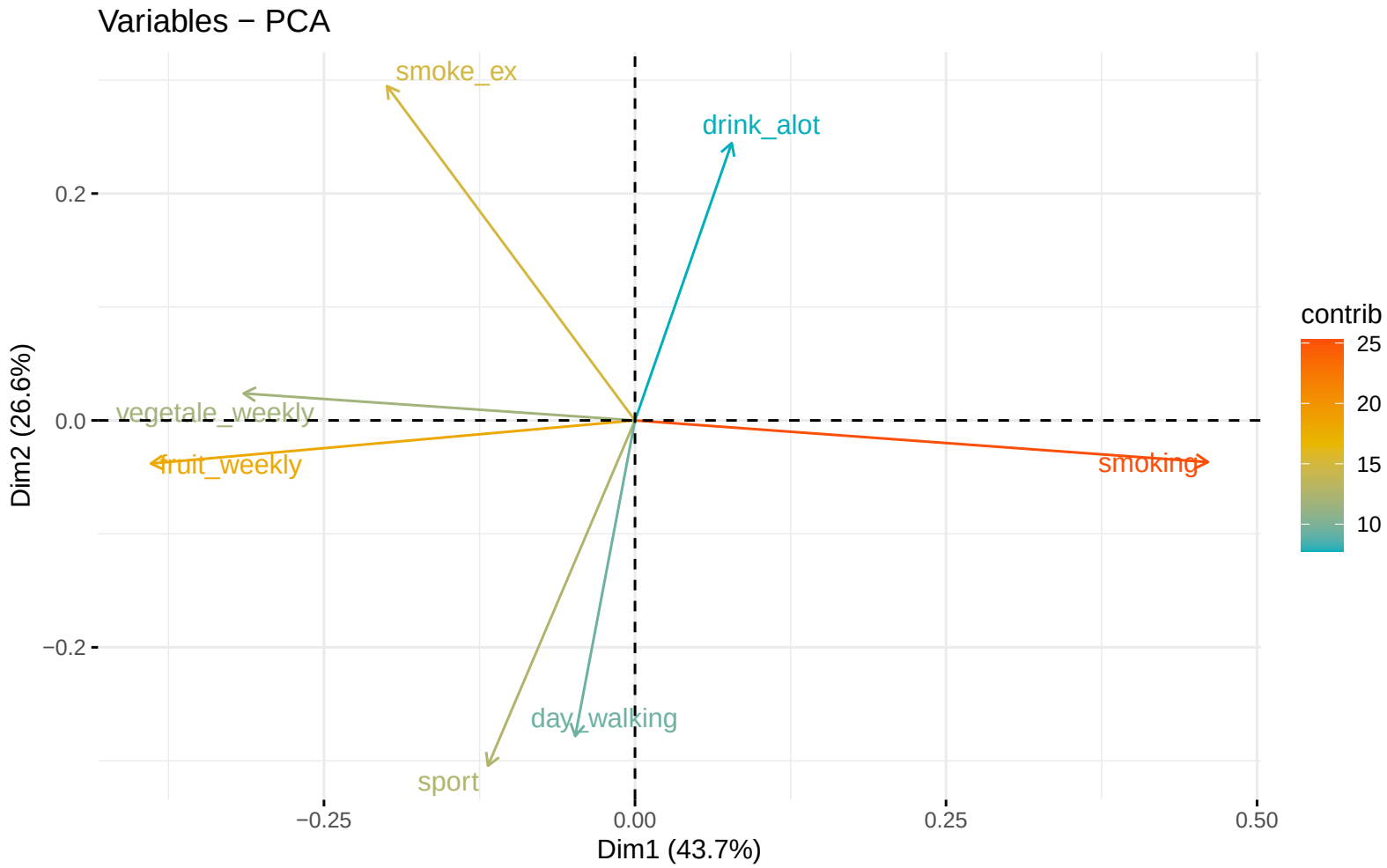
- Introduced by Zeileis et al. (2008): first estimate a model in the population:

$$\text{e.g. : } h_i = \beta_0 + \beta_1 \times \text{LifeStyle}_i$$

- then perform a sequence of test on the instability of the parameters across possible subgroups;

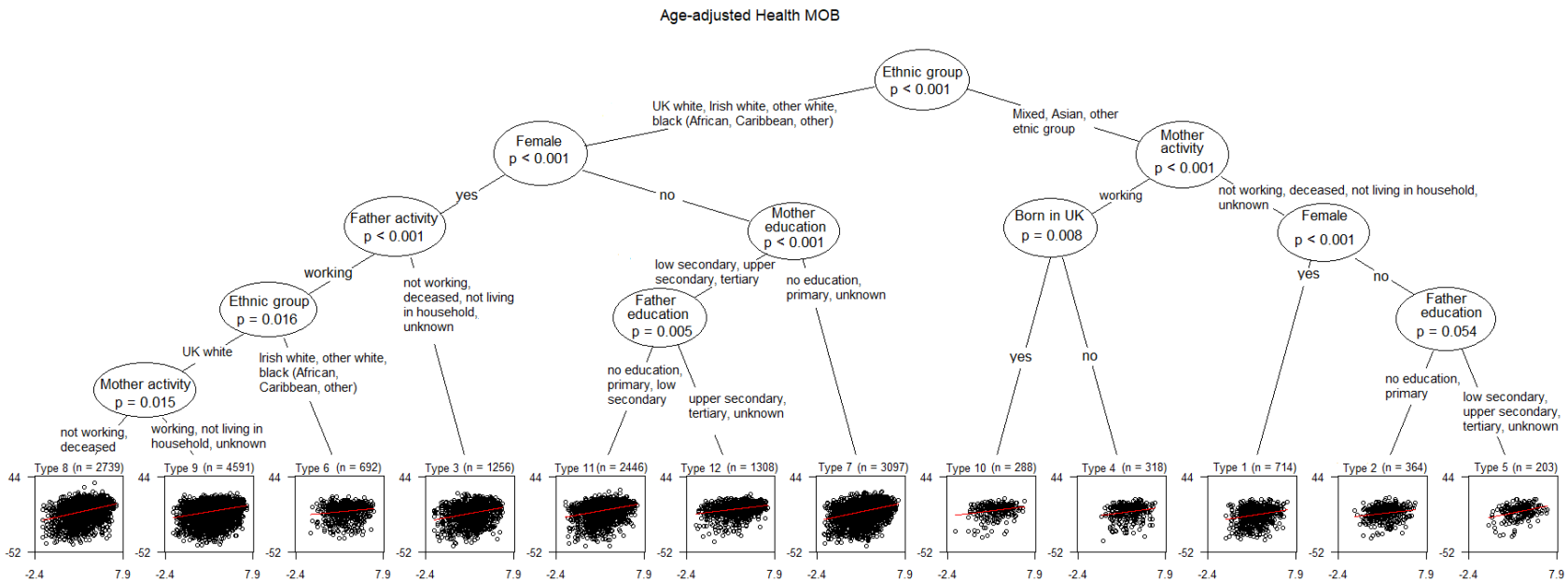
- stop when a further split does not improve the out-of-sample prediction accuracy.

Health inequality of opportunity in UK: choices



source: Brunori, Davillas, Jones, Scarchilli (2021) on UKHLS 2010-16

Health inequality of opportunity in UK: MOB



source: Brunori, Davillas, Jones, Scarchilli (2021) on UKHLS 2010-16

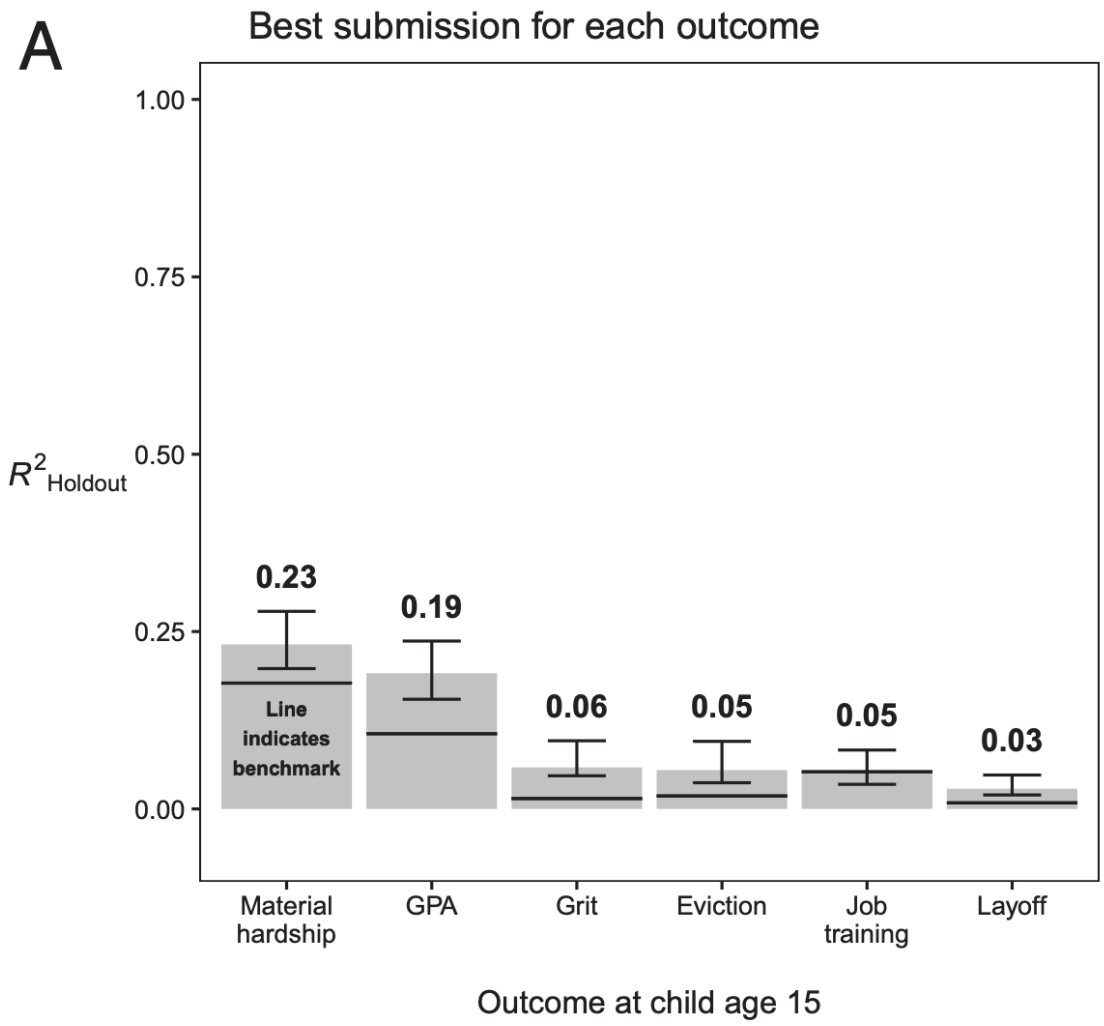
Conclusions

- Interpreting IOP as a prediction problem may deepen our understanding of inequality (e.g. intersectionality);

- tree-based methods are not particularly sophisticated but they well handle the ML 'accuracy-interpretability' trade-off;

- evaluating the predictive ability of our model tells us a single fundamental truth: we know very little!

Fragile Families Challenge



source: Salganik et al. (2020)