Tree-based models to estimate inequality of opportunity

Sapienza Economics Seminars
09.04.2021

Paolo Brunori
University of Florence and University of Bari

This presentation

Should machine learning replace traditional econometric methods when estimating inequality of opportunity?

Based on material co-authored with:

Paul Hufe, Daniel Mahler, and Guido Neidöfer.

Motivation

- formal definitions by political philosophers: Rawls (1971), Dworkin (1981), Arneson (1989), Cohen (1989)
- economics: Roemer (1998), Fleurbaey (2008)
- Roemer's model triggered a still developing literature on the measurement of inequality of opportunity (IOP)
- can supervised machine learning (tree-based algorithms) contribute?

Roemer's Model

$$y_i = f(\mathbf{C}_i, e_i) + u_i$$

- y_i: individual's i outcome;
- $\mathbf{C}_i$: circumstances beyond individual control;
- e_i: effort;
- u_i: random component.

Romer's society

$\mathbf{c}_1$	$\mathbf{c}_2$	c	$e = L$	$e = M$	$e = H$
W	$\mathbf{M}$	1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$
W	F	2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$
B	M	3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$
B	F	4	$y_{4,1}$	$y_{4,2}$	$y_{4,3}$

Romer's society

$\mathbf{c}_1$	$\mathbf{c}_2$	c	$e = L$	$e = M$	$e = H$
W	**M**	1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$
W	F	2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$
B	M	3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$
B	F	4	$y_{4,1}$	$y_{4,2}$	$y_{4,3}$

TYPE

TRANCHE

Ex-post EOP (Roemer, 1998)

- ignoring the random component:

$$e_i = e_j \cap C_i = C_j \to y_i = y_j \ , \ \ \forall i, j \in 1, ..., n$$

- equality of opportunity is satisfied if:

$$e_i = e_j \to y_i = y_j \ , \ \ \forall i, j \in 1, ..., n$$

$\Rightarrow$ IOP = within-tranche inequality (ex-post)

Ex-ante EOP (Van de gaer, 1993)

- EOP = equality of opportunity sets' value;
- opportunity set = type-specific outcome distribution;
- utilitarian value = types' outcome mean;

$\Rightarrow$ IOP = between-type inequality (ex-ante).

Implementation

- ex-ante:
  - identify types;
  - estimate between-type inequality.
- ex-post:
  - identify types;
  - identify tranches;
  - estimate within-tranche inequality.

Empirical approaches (ex-ante)

- two empirical approaches:
    - non-parametric (Checchi and Peragine, 2010);
    - parametric (Ferreira and Gignoux, 2011).

Non-parametric approach

- exactly implement the idea of Romerian types;
- data-hungry: sex, race, parents' ISCO 1 digit = 400 types;
- sparsely populated types → large measurement error;
- measurement error → upward biased inequality (Chakravarty and Eichhron, 1994);
- *ad hoc* solutions → downward bias.

Parametric approach

- reduced form model: $\ln y_i = \beta \mathbf{C}_i + u_i$;

- $IOP = I(\hat{y})$;

- additive and log linear functional form → downward bias.

Problems with existing methods

- arbitrary use of information: Checchi et al. (2016) 92 types, Suárez and Menéndez (2017) 360 types;
- arbitrary assumptions on $f()$;
- **C** partial observability → downward bias:
  - rich dataset (Biörklund et al., 2012; Hufe et al., 2017);
  - interactions (Hufe and Peichl, 2015).
- back to risk of upward biased IOP.

IOP as a prediction problem

- how predictive of y are observable $\mathbf{C}$?

- unknown data generating process;

- a meaningless exercize “in sample”;

- a prediction problem “out-of-sample”.

ML - IOP analogy

- $I\hat{O}P = IOP + \downarrow$ (unobserved $\mathbf{C}$/restrictions)$+ \uparrow$ (variance);

- $\mathbb{E}[(\hat{y} - y)]^2 = bias^2 + variance + \epsilon^2$;

- ML: choose the model that minimizes out-of-sample error;

- IOP: choose the model that maximizes out-of-sample IOP.

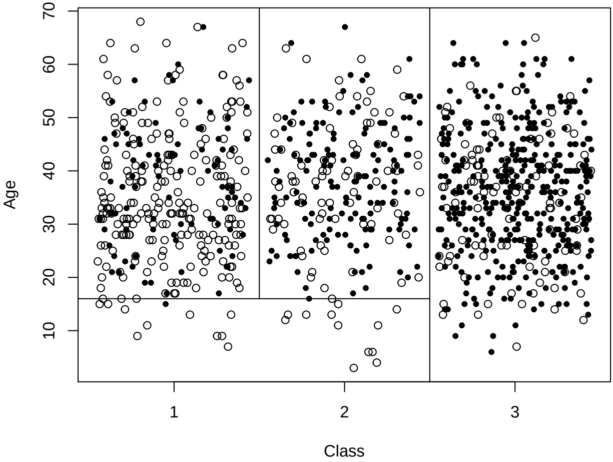
Tree-based algorithms

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

- the population is divided into non-overlapping subgroups;

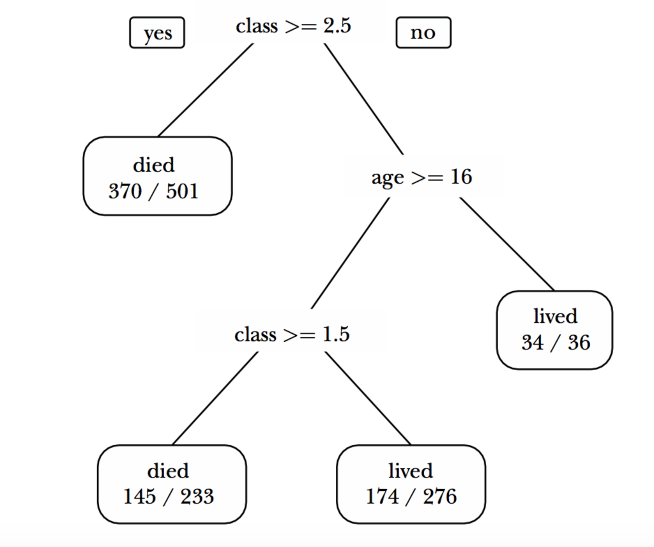
- prediction of each observation is the the mean value of the dependent variable in the group.

What is a tree? cnt.



Source: adapted from Varian, 2014

What is a tree? cnt.



Source: Varian, 2014

What is a tree? cnt.

- deep trees fit well “in-sample”;
- but perform poorly out-of-sample;
- different solutions lead to different type of trees;
- we use *conditional inference trees* (Hothorn et al., 2006).

Conditional inference trees

- test the null hypothesis of independence, $H^{C_p} = D(Y|Cp) = D(Y), \forall C_p \in \mathbf{C}$;
- no (adjusted) p-value $< \alpha \rightarrow$ exit the algorithm;
- select the variable, $C^\star$, with the lowest $p-$value;
- test the discrepancy between the subsamples for each possible binary partition based on $C^\star$;
- split the sample by selecting the splitting point that yields the lowest p-value;
- repeat the algorithm for each of the resulting subsamples.

Opportunity trees: *pros*

- the selection of $\mathbf{C}$ is no longer arbitrary;
- $\hat{f}()$ becomes endogenous to data;
- provide a test for the null hypothesis of EOP;
- tell a story about the opportunity structure.

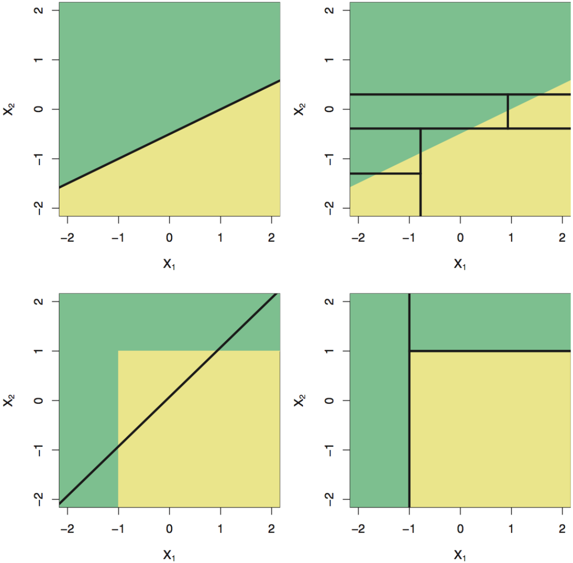
Opportunity trees: *cons*

- misleading when two or more controls are highly correlated;

- perform poorly if the data generating process is linear;

- sensitive to the specific sample observed.

DGP linearity

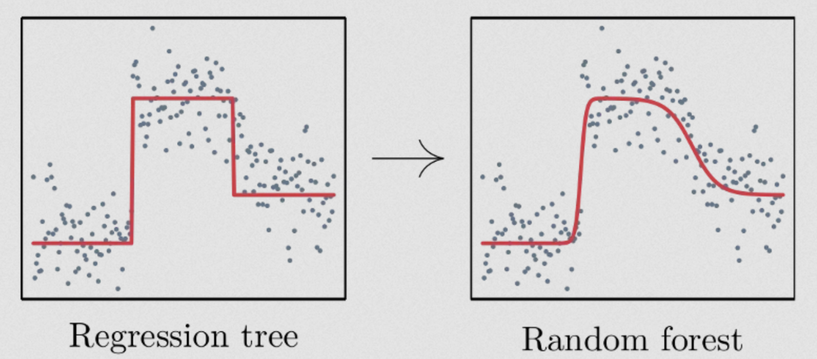


source: James et al. (2013)

Random forests

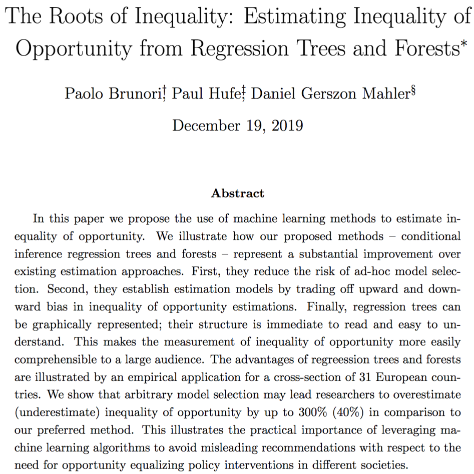
- random forests improve tree's predictive performance;
- a forest is made of hundreds of conditional inference trees;
- each tree uses a subsample of observations and each split a subsample of regressors.

Trees and forests



source: Schlosser et al. (2018)

Method validation



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

Paolo Brunori[†] Paul Hufe[‡] Daniel Gerszon Mahler[§]

December 19, 2019

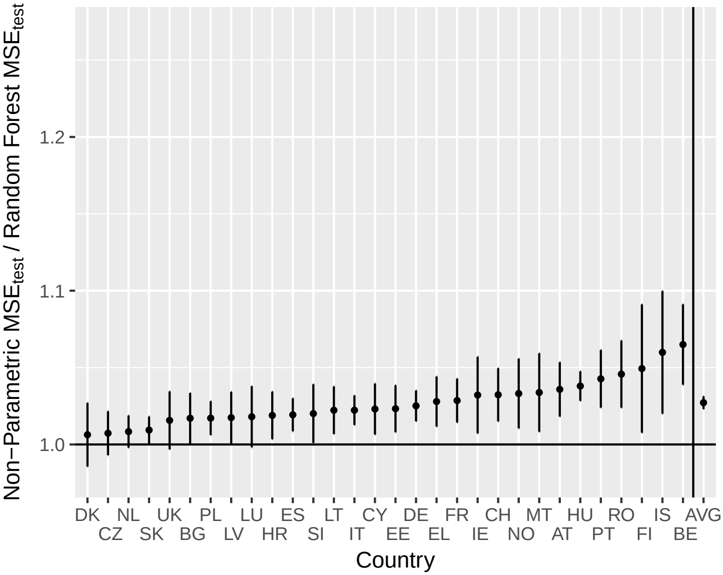
Abstract

In this paper we propose the use of machine learning methods to estimate inequality of opportunity. We illustrate how our proposed methods – conditional inference regression trees and forests – represent a substantial improvement over existing estimation approaches. First, they reduce the risk of ad-hoc model selection. Second, they establish estimation models by trading off upward and downward bias in inequality of opportunity estimations. Finally, regression trees can be graphically represented; their structure is immediate to read and easy to understand. This makes the measurement of inequality of opportunity more easily comprehensible to a large audience. The advantages of regreession trees and forests are illustrated by an empirical application for a cross-section of 31 European countries. We show that arbitrary model selection may lead researchers to overestimate (underestimate) inequality of opportunity by up to 300% (40%) in comparison to our preferred method. This illustrates the practical importance of leveraging machine learning algorithms to avoid misleading recommendations with respect to the need for opportunity equalizing policy interventions in different societies.

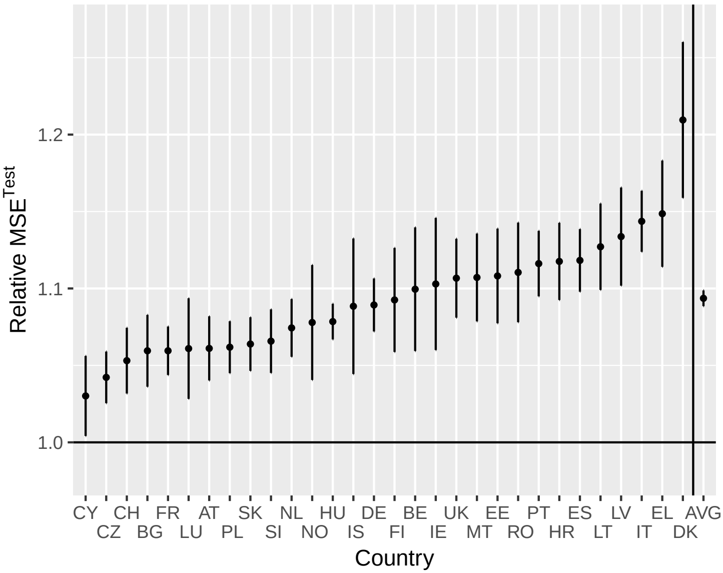
Relative performance

- conditional inference regression trees and forests
- compared with traditional approaches:
  - □ non-parametric approach 40 types (Checchi et al., 2016);
  - □ parametric approach 20 regressors (Palomino et al., 2016);
- 31 countries (EU-SILC, 2011).

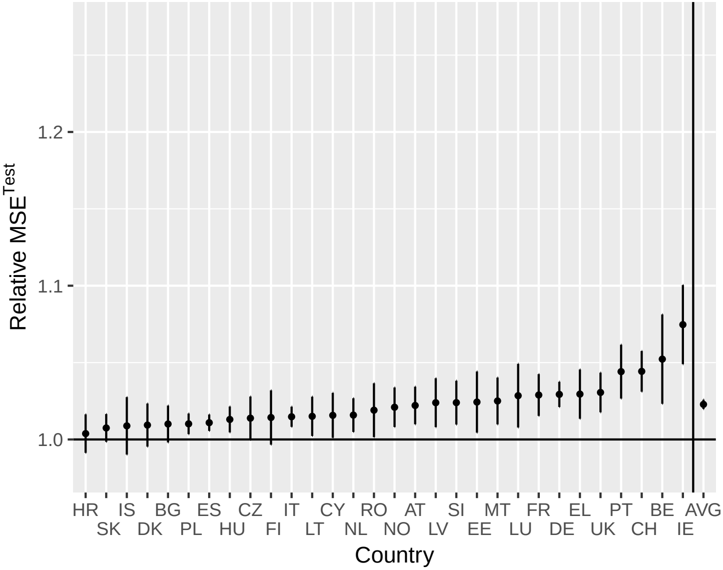
Random forest vs. non-parametric



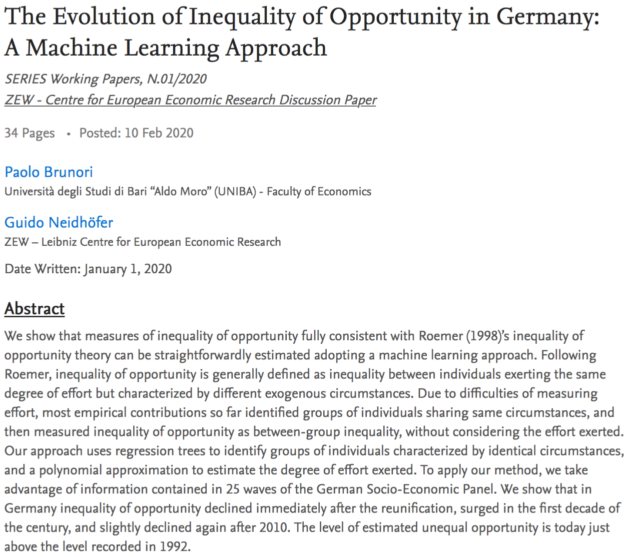
Random forest vs. parametric



Random forest vs. conditional inference trees



Ex-post IOP



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

SERIES Working Papers, N.01/2020
ZEW - Centre for European Economic Research Discussion Paper

34 Pages • Posted: 10 Feb 2020

Paolo Brunori
Università degli Studi di Bari "Aldo Moro" (UNIBA) - Faculty of Economics

Guido Neidhöfer
ZEW – Leibniz Centre for European Economic Research

Date Written: January 1, 2020

Abstract

We show that measures of inequality of opportunity fully consistent with Roemer (1998)'s inequality of opportunity theory can be straightforwardly estimated adopting a machine learning approach. Following Roemer, inequality of opportunity is generally defined as inequality between individuals exerting the same degree of effort but characterized by different exogenous circumstances. Due to difficulties of measuring effort, most empirical contributions so far identified groups of individuals sharing same circumstances, and then measured inequality of opportunity as between-group inequality, without considering the effort exerted. Our approach uses regression trees to identify groups of individuals characterized by identical circumstances, and a polynomial approximation to estimate the degree of effort exerted. To apply our method, we take advantage of information contained in 25 waves of the German Socio-Economic Panel. We show that in Germany inequality of opportunity declined immediately after the reunification, surged in the first decade of the century, and slightly declined again after 2010. The level of estimated unequal opportunity is today just above the level recorded in 1992.

Ex-post IOP

- Trees identify types;
- a precondition to estimate IOP *á la* Roemer (ex-post);
- main challenge: effort identification.

Effort

- Roemer's identification strategy, two assumptions:

  1. orthogonality: $e \perp\!\!\!\perp C$

  2. monotonicity: $\frac{\partial f}{\partial e} \geq 0$

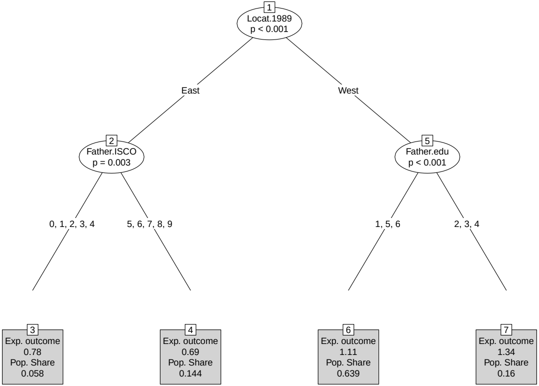
Degree of effort

- with observable effort = quantile of the type-specific effort distribution;
- with unobservable effort = quantile of the type-specific outcome distribution;
- our approach: (Bernstein) polynomial approximation of the type-specific outcome distribution.

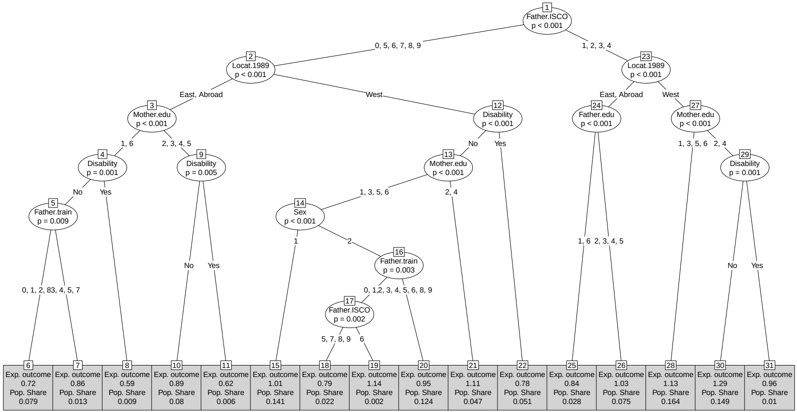
Data

- SOEP (v33) including all subsamples apart from the refugee samples;
- 25 waves 1992-2016;
- adult individuals (30-60);
- circumstances considered: migration background, location in 1989, mother's education, father's education, father's occupation, father's training, month of birth, disability, siblings;
- outcome: 'age-adjusted' household equivalized income.

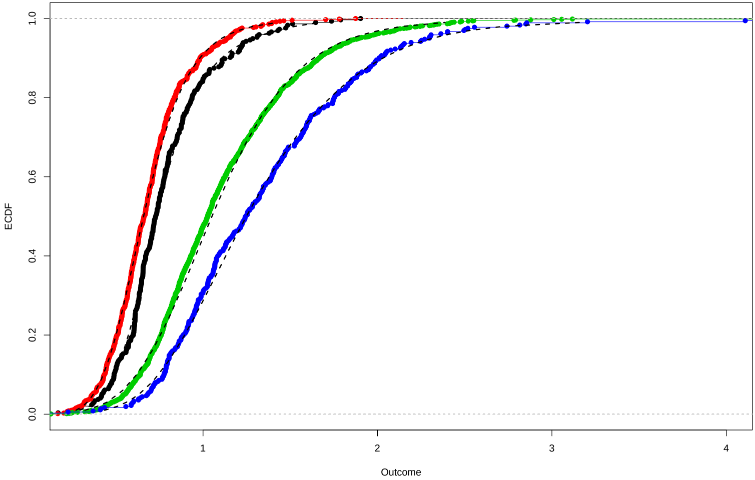
Opportunity tree in 1992



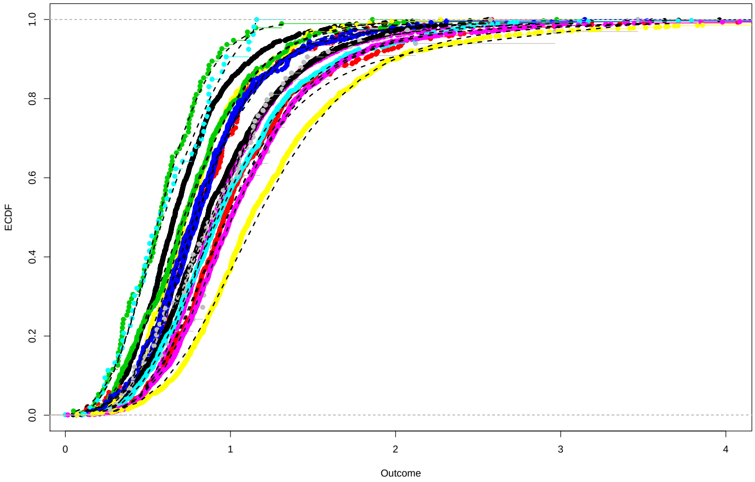
Opportunity tree in 2016



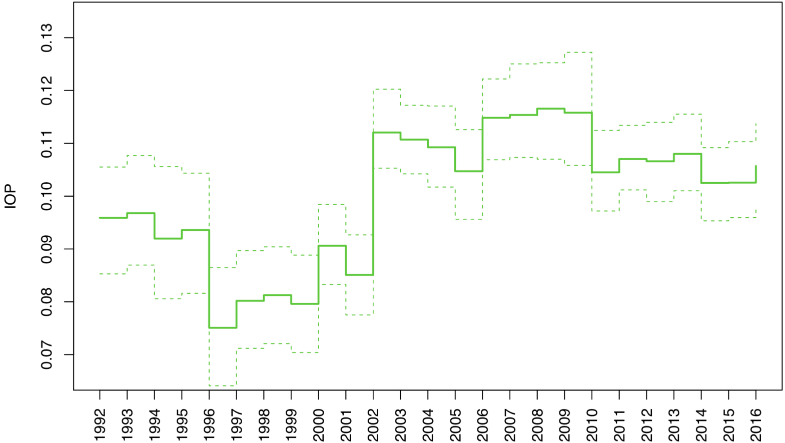
IOP in 1992



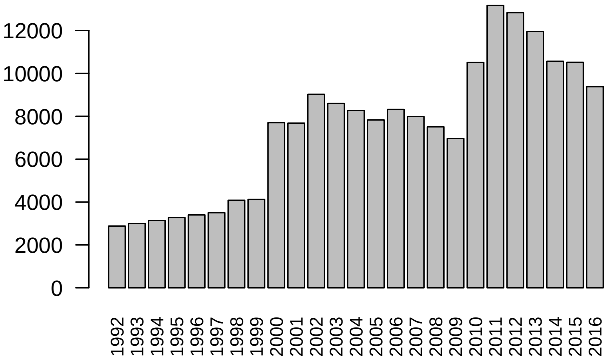
IOP in 2016



IOP trend 1992-2016

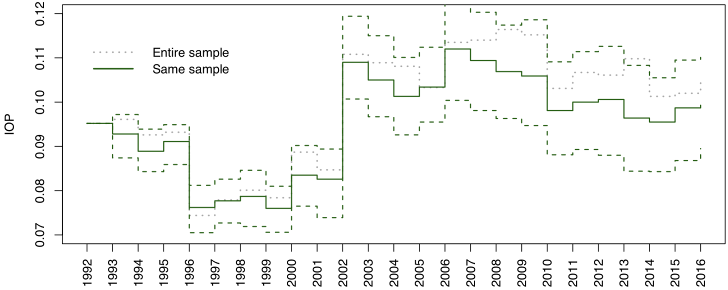


Sample size 1992-2016

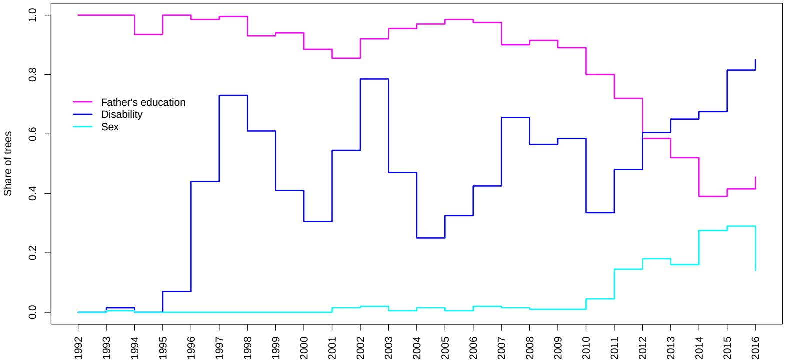


Mean IOP trend 1992-2016 (same sample size)

(b) IOP



Circumstances' relative importance

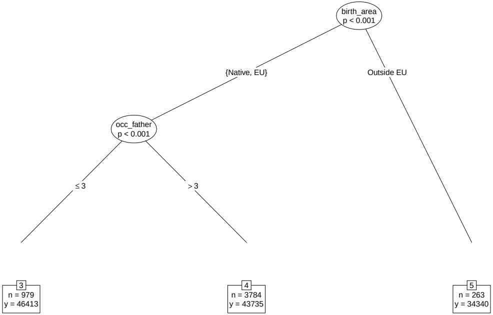


Conclusions

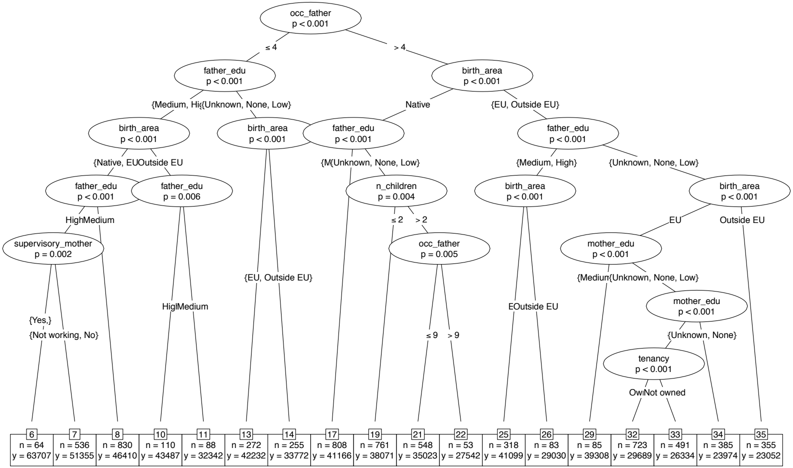
- many other ML approaches can be used:
  - unsupervised learning such as Li Donni et al. (2015) and Carrieri and Jones (2020)
  - best subset regression (EqualChances.org)
  - LASSO (or other regularization methods) as for example Hufe et al. (2019)
- but there exists a second key trade off in ML: complexity Vs. interpretability.

Additional material trees EU-SILC

Norway

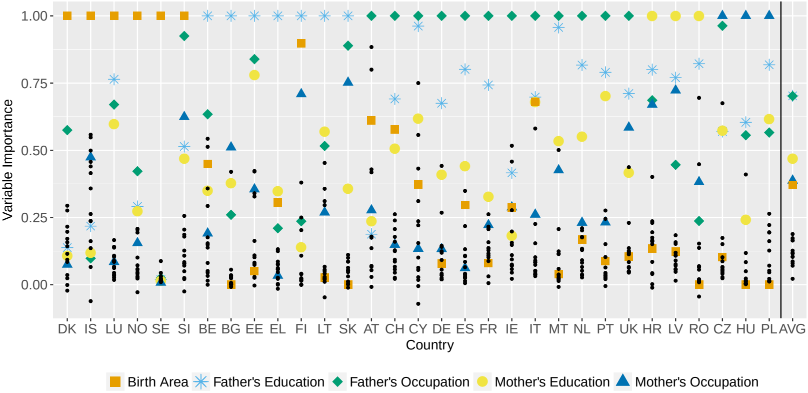

birth_area
p < 0.001
{Native, EU}
Outside EU
occ_father
p < 0.001
≤ 3
> 3
3
n = 979
y = 46413
4
n = 3784
y = 43735
5
n = 263
y = 34340

Luxembourg

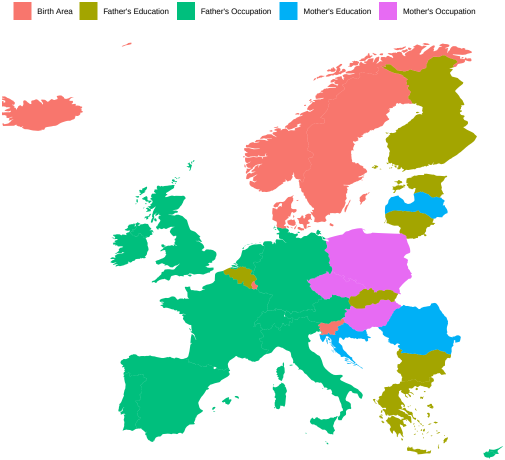

occ_father
p < 0.001
≤ 4
> 4
father_edu
p < 0.001
birth_area
p < 0.001
{Medium, Hi{Unknown, None, Low}
Native
{EU, Outside EU}
birth_area
p < 0.001
birth_area
p < 0.001
father_edu
p < 0.001
father_edu
p < 0.001
{Native, EUOutside EU
{M{Unknown, None, Low}
{Medium, High}
{Unknown, None, Low}
father_edu
p < 0.001
father_edu
p = 0.006
n_children
p = 0.004
birth_area
p < 0.001
birth_area
p < 0.001
HighMedium
≤ 2
> 2
EU
Outside EU
supervisory_mother
p = 0.002
occ_father
p = 0.005
mother_edu
p < 0.001
{EU, Outside EU}
{Mediun{Unknown, None, Low}
HiglMedium
EOutside EU
mother_edu
p < 0.001
{Yes,}
{Not working, No}
≤ 9
> 9
{Unknown, None}
tenancy
p < 0.001
OwiNot owned
6
n = 64
y = 63707
7
n = 536
y = 51355
8
n = 830
y = 46410
10
n = 110
y = 43487
11
n = 88
y = 32342
13
n = 272
y = 42232
14
n = 255
y = 33772
17
n = 808
y = 41166
19
n = 761
y = 38071
21
n = 548
y = 35023
22
n = 53
y = 27542
25
n = 318
y = 41099
26
n = 83
y = 29030
29
n = 85
y = 39308
32
n = 723
y = 29689
33
n = 491
y = 26334
34
n = 385
y = 23974
35
n = 355
y = 23052

Additional material
variables importance EU-SILC

Variables importance

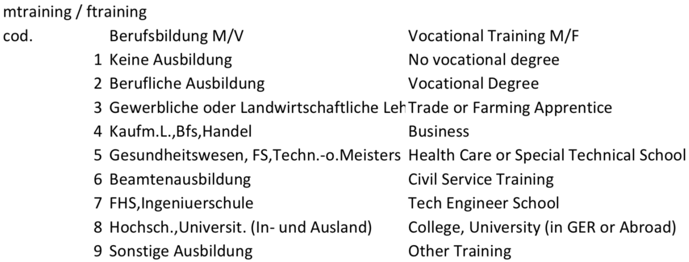


Variables importance, cnt



Additional material: training/education definition SOEP

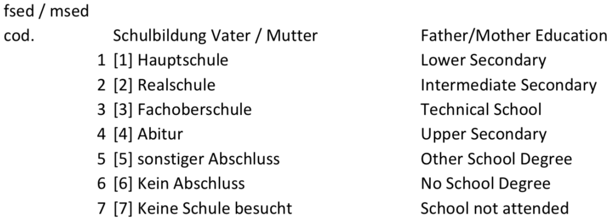
Mother/father raining



mtraining / ftraining

cod.		Berufsbildung M/V	Vocational Training M/F
	1	Keine Ausbildung	No vocational degree
	2	Berufliche Ausbildung	Vocational Degree
	3	Gewerbliche oder Landwirtschaftliche Leh	Trade or Farming Apprentice
	4	Kaufm.L.,Bfs,Handel	Business
	5	Gesundheitswesen, FS,Techn.-o.Meisters	Health Care or Special Technical School
	6	Beamtenausbildung	Civil Service Training
	7	FHS,Ingeniuerschule	Tech Engineer School
	8	Hochsch.,Universit. (In- und Ausland)	College, University (in GER or Abroad)
	9	Sonstige Ausbildung	Other Training

Mother/father education



fsed / msed			
cod.		Schulbildung Vater / Mutter	Father/Mother Education
	1	[1] Hauptschule	Lower Secondary
	2	[2] Realschule	Intermediate Secondary
	3	[3] Fachoberschule	Technical School
	4	[4] Abitur	Upper Secondary
	5	[5] sonstiger Abschluss	Other School Degree
	6	[6] Kein Abschluss	No School Degree
	7	[7] Keine Schule besucht	School not attended