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

Conference on Inequality of Opportunity
University of Queensland & Griffith University
27-28 June, 2019

Paolo Brunori
University of Florence & University of Bari

- "The Roots of Inequality Estimating Inequality of Opportunity from Regression Trees and Forests" is a joint work with Paul Hufe (Ifo) and Daniel G. Mahler (World Bank);

Ex-ante IOP estimation

- $y = g(C) + \epsilon$
- causality is conceptually and empirically excluded $\rightarrow$ covariance of the outcome and circumstances' variability

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

I is a suitable inequality index;

$\hat{y}$ is the predicted outcome distribution based on $\hat{g}(C)$;

$\hat{g}()$ is estimated on survey data.

Typical machine learning domain

- unknown data generating process;
- need to establish a reliable empirical link between a set of controls and an outcome.

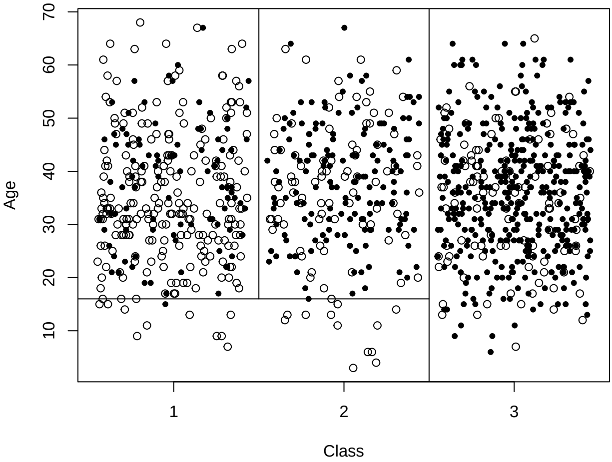
ML and IOP

- ML: bias-variance trade-off;
- IOP partial observability (downward bias) - sampling variance (upward bias);
- ML: choose the model that minimizes out-of-sample MSE;
- IOP: choose the model that maximizes IOP out-of-sample (Social Choice and Welfare - 2019 with Peragine and Serlenga).

Trees

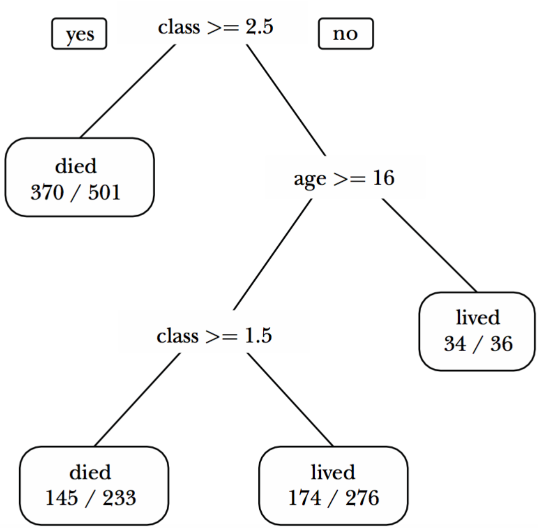
- among supervised learning algorithms regression trees seem an obvious choice;
- a tree is an algorithm to predict a dependent variable based on observable predictors (Morgan and Sonquist,1963; Breiman et al.,1984);
- the population is divided into non-overlapping subgroups based on a partition of the 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

Tuning

- 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 statistical test (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.

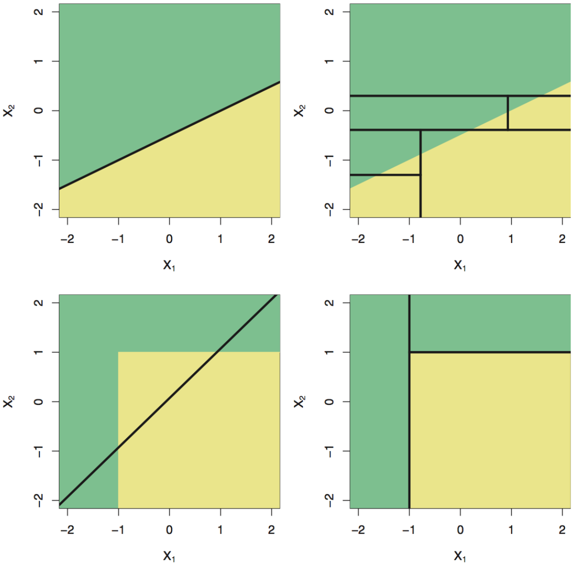
Opportunity trees: *pros*

- the selection of $\mathbf{C}$ is no longer arbitrary;
- the model specification is 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.



source: James et al. (2013)

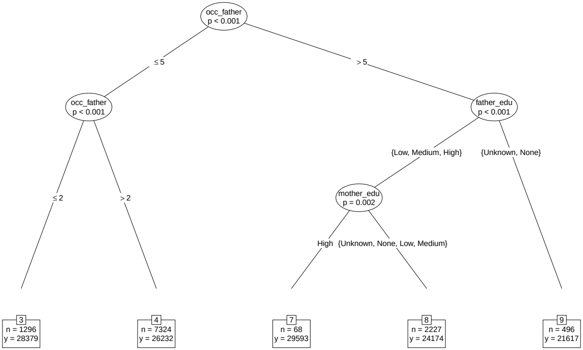
Random forests

- 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, at each splitting point, a subsample of controls.

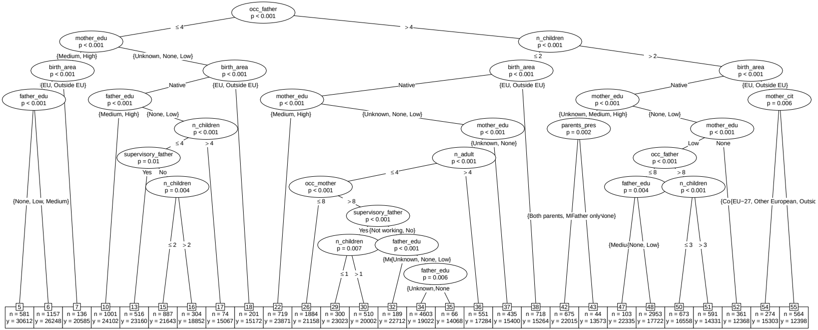
data

- EU-SILC 2011;
- subsample: adults (30-60);
- y: household equivalent disposable income;
- **C**: 21 questions about respondents' background (sex, birth area, proxies for socioeconomic status);
- already used to estimate IOP.

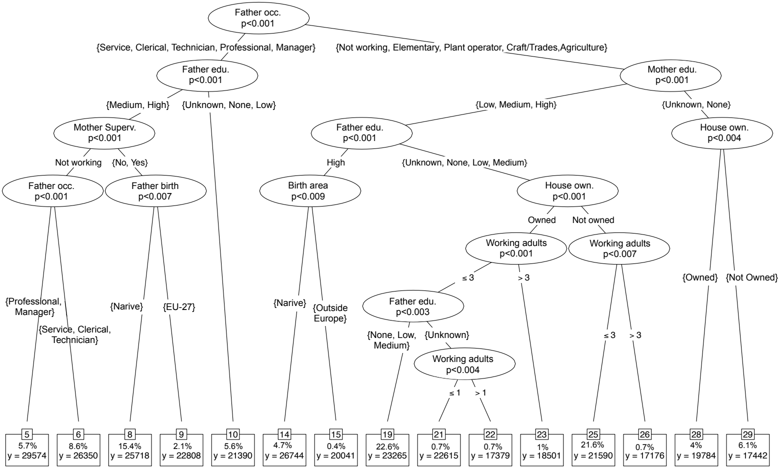
The Netherlands



Italy


occ_father p < 0.001
≤ 4
> 4
mother_edu p < 0.001
{Medium, High}
{Unknown, None, Low}
n_children p < 0.001
≤ 2
> 2
birth_area p < 0.001
{EU, Outside EU}
Native
father_edu p < 0.001
{None, Low, Medium}
{Medium, High}
{None, Low}
n_children p < 0.001
supervisory_father p = 0.01
Yes
No
n_children p = 0.004
mother_edu p < 0.001
{Unknown, None, Low}
mother_edu p < 0.001
{Unknown, None}
n_adult p < 0.001
occ_mother p < 0.001
≤ 8
> 8
supervisory_father p < 0.001
{Not working, No}
n_children p = 0.007
≤ 1
> 1
father_edu p < 0.001
{Unknown, None, Low}
father_edu p = 0.006
{Unknown,None
{Unknown, Medium, High}
parents_pres p = 0.002
{Both parents, M
Father only
mother_edu p < 0.001
Low
None
occ_father p < 0.001
father_edu p = 0.004
{None, Low}
n_children p < 0.001
≤ 3
> 3
mother_cit p = 0.006
{EU−27, Other European, Outsic
5 n = 581 y = 30612
6 n = 1157 y = 26248
7 n = 136 y = 20585
10 n = 1001 y = 24102
13 n = 516 y = 23160
15 n = 887 y = 21643
16 n = 304 y = 18852
17 n = 74 y = 15067
18 n = 201 y = 15172
22 n = 719 y = 23871
26 n = 1884 y = 21158
29 n = 300 y = 23023
30 n = 510 y = 20002
32 n = 189 y = 22712
34 n = 4603 y = 19022
35 n = 66 y = 14068
36 n = 551 y = 17284
37 n = 435 y = 15400
38 n = 718 y = 15264
42 n = 675 y = 22015
43 n = 44 y = 13573
47 n = 103 y = 22335
48 n = 2953 y = 17722
50 n = 673 y = 16558
51 n = 591 y = 14331
52 n = 361 y = 12368
54 n = 274 y = 15303
55 n = 564 y = 12398

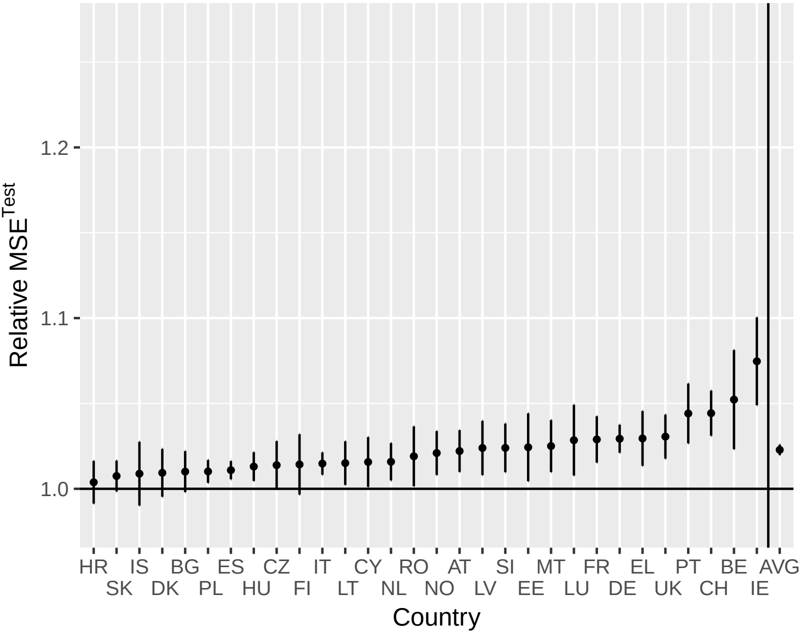
Germany


Father occ.
p<0.001
{Service, Clerical, Technician, Professional, Manager}
{Not working, Elementary, Plant operator, Craft/Trades,Agriculture}
Father edu.
p<0.001
Mother edu.
p<0.001
{Medium, High}
{Unknown, None, Low}
{Low, Medium, High}
{Unknown, None}
Mother Superv.
p<0.001
Father edu.
p<0.001
House own.
p<0.004
Not working
{No, Yes}
High
{Unknown, None, Low, Medium}
Father occ.
p<0.001
Father birth
p<0.007
Birth area
p<0.009
House own.
p<0.001
Owned
Not owned
Working adults
p<0.001
Working adults
p<0.007
≤ 3
> 3
{Owned}
{Not Owned}
{Professional,
Manager}
{Service, Clerical,
Technician}
{Narive}
{EU-27}
{Narive}
{Outside
Europe}
Father edu.
p<0.003
{None, Low,
Medium}
{Unknown}
≤ 3
> 3
Working adults
p<0.004
≤ 1
> 1
5
5.7%
y = 29574
6
8.6%
y = 26350
8
15.4%
y = 25718
9
2.1%
y = 22808
10
5.6%
y = 21390
14
4.7%
y = 26744
15
0.4%
y = 20041
19
22.6%
y = 23265
21
0.7%
y = 22615
22
0.7%
y = 17379
23
1%
y = 18501
25
21.6%
y = 21590
26
0.7%
y = 17176
28
4%
y = 19784
29
6.1%
y = 17442

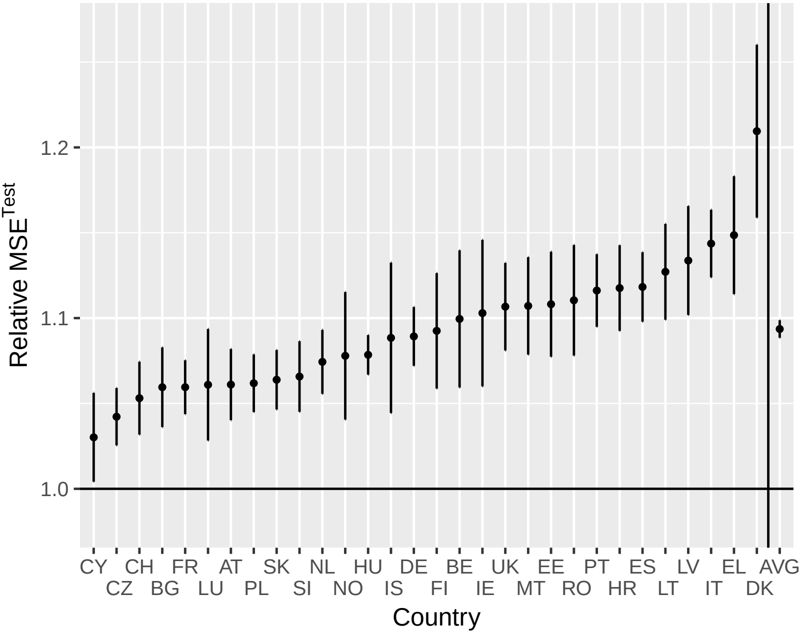
Random forests

- random forests of 200 conditional inference trees used to:
  - estimate IOp;
  - quantify relative variable importance.

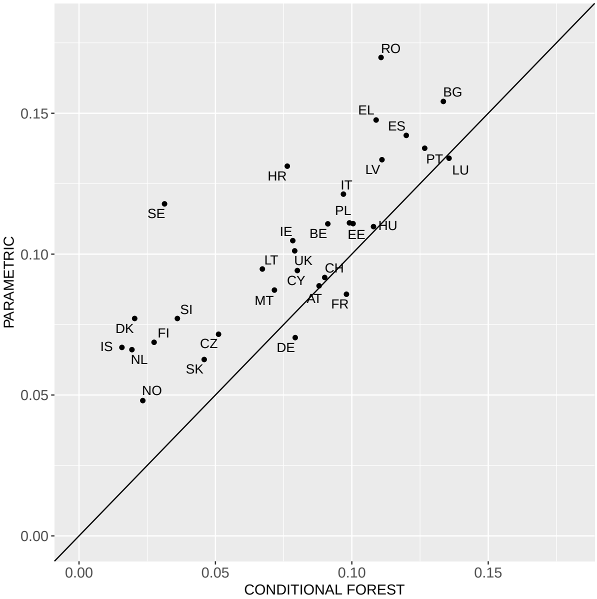
Predictive performance: trees Vs. forest



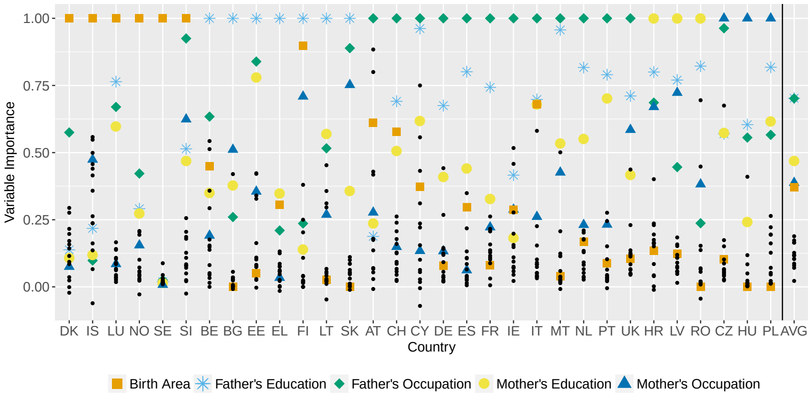
Predictive performance: Parametric Vs. forest



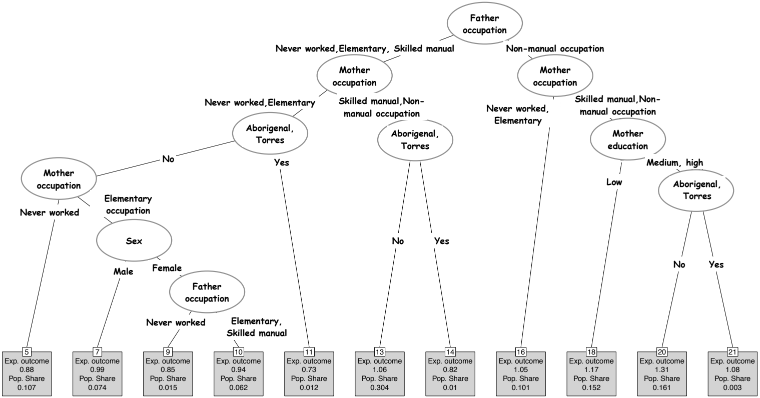
Estimates: Parametric Vs. forest



Variables importance



Bonus tree: Australia, 2015



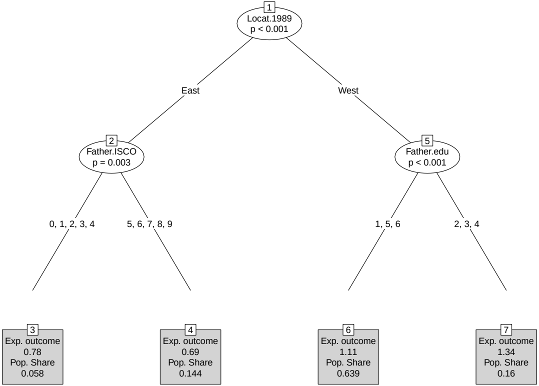
Ex-post IOP
(joint work in progress with Guido Neidhöfer)

- Conditional inference regression trees have two important advantage
  - they identify types;
  - they are parsimonious.
- having types with sufficient sample size one can move further and estimate IOP consistently with Roemer's original theory;
- ex-post IOP definition is based on the estimation of the type-specific outcome distribution.

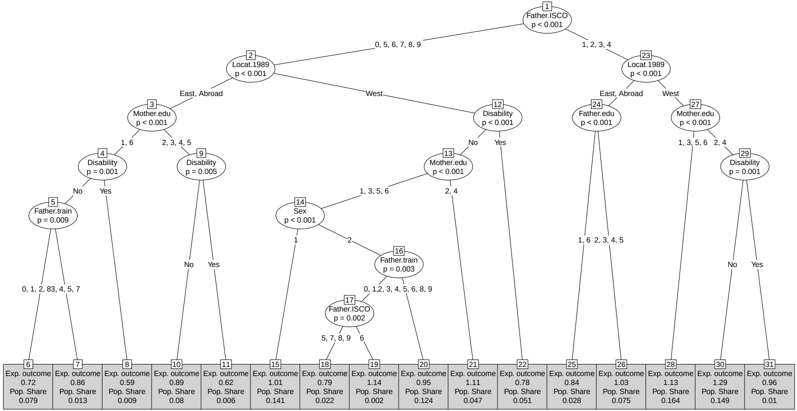
Effort

- According to Roemer the quantile of the type-specific outcome distribution is a convincing proxy of the degree of effort exerted;
- ex-post IOP quantifies to what extent individuals exerting the same degree of effort do not obtain the same outcome;
- we use Bernstain polynomial approximation of the types' ECDF to measure ex-post IOP.

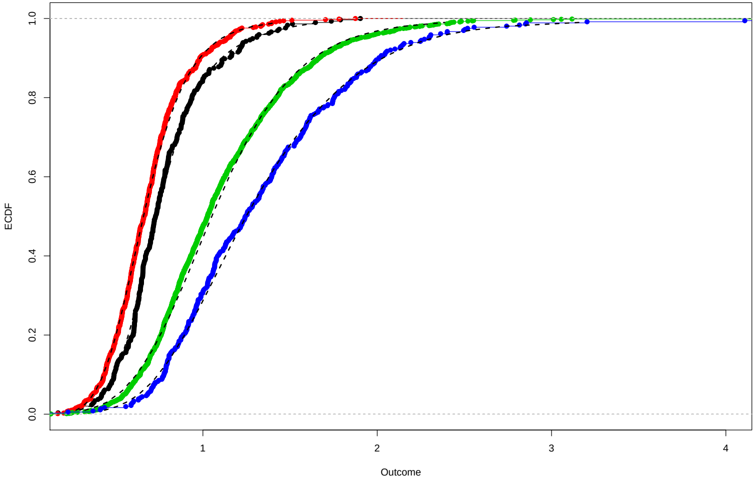
Opportunity tree in 1992



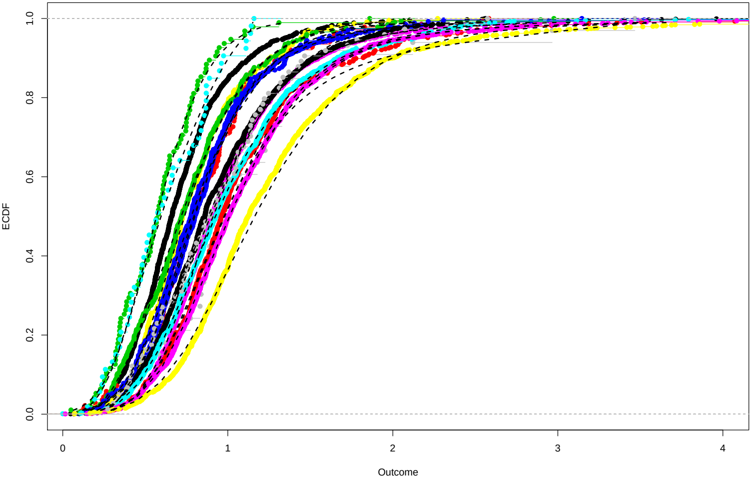
Opportunity tree in 2016



IOP in 1992



IOP in 2016

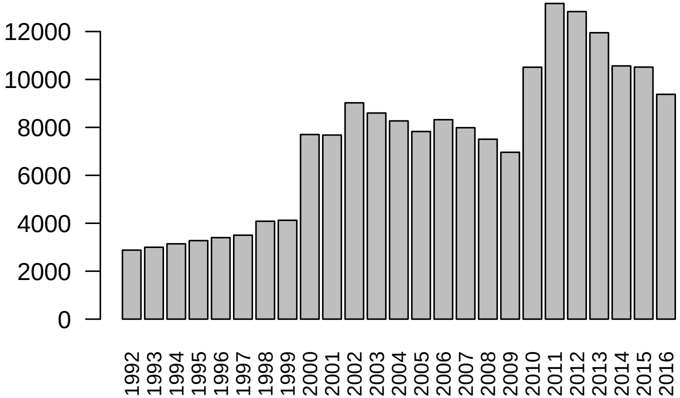


Conclusions

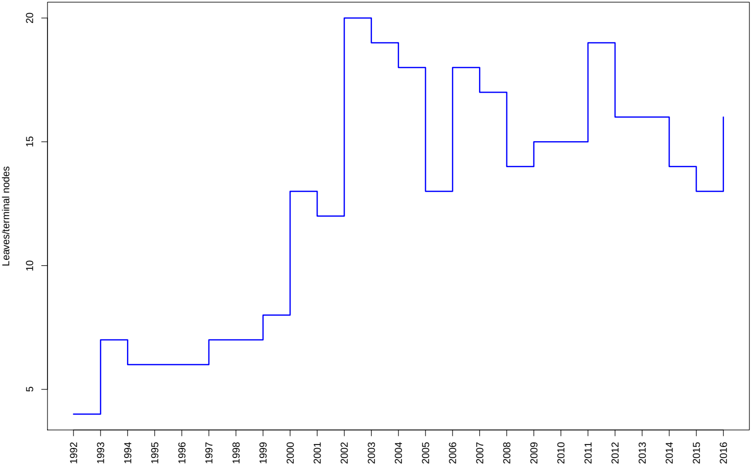
- many other ML approaches can be used:
  - unsupervised learning such as Li Donni et al. (2015) and Wu, Trivedi, Rao, Tang (2018)
  - best subset regression (EqualChances.org)
  - LASSO (or other regularization methods) as for example Hufe et al. (2019)
- but there exists a second key trade off in ML: complexity Vs. interpretability.

Additional material: trend in Germany

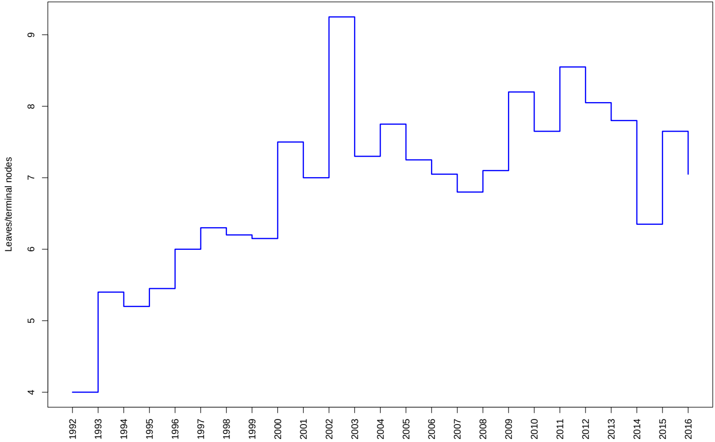
Sample size 1992-2016



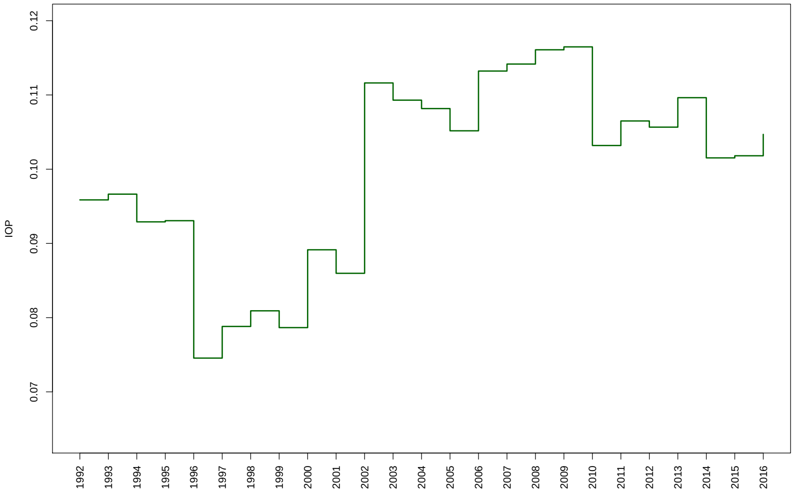
Number of types 1992-2016



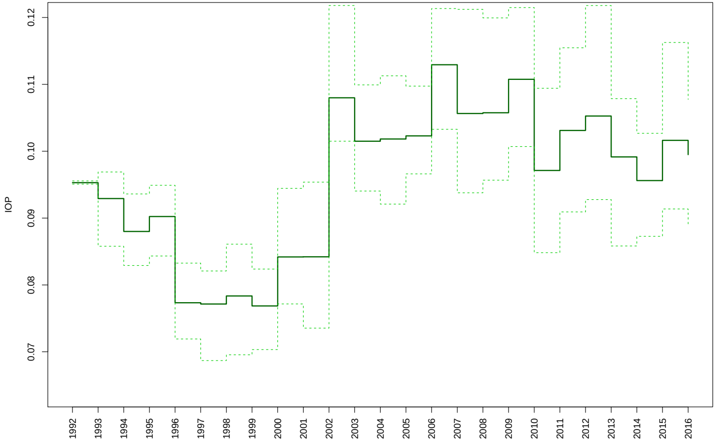
Mean number of types (same sample size) 1992-2016



IOP trend 1992-2016



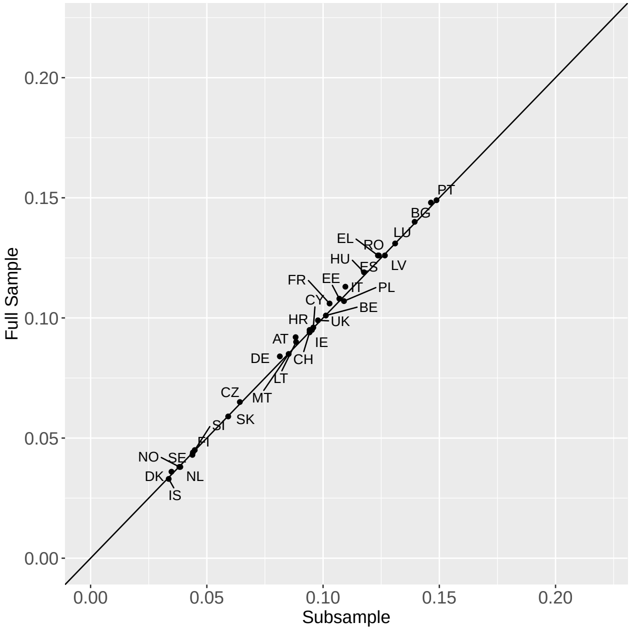
Mean IOP trend 1992-2016 (same sample size)



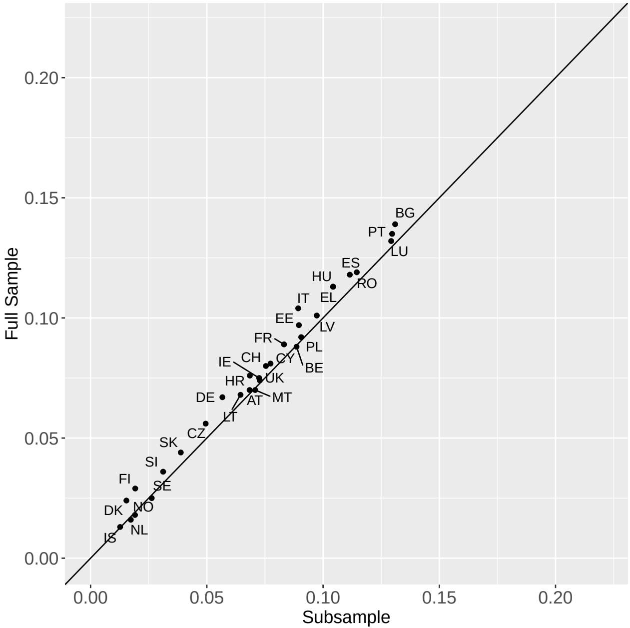
Confidence bounds are the 0.975 and 0.025 quantiles of the distribution of IOP estimates.

Additional material: sample size EU-SILC

Sensitivity to sample size: forests



Sensitivity to sample size: trees



Sensitivity to sample size: parametric