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

Inequality and Opportunities
Universidad Complutense de Madrid
24.05.2019

Paolo Brunori
University of Florence

Guido Neidhöfer
ZEW

EOp

- equality of opportunity (EOP): a very successful political ideal
- two reasons:
  1. EOP = equality + freedom
  2. EOP is sufficiently vague

Literature

- a “third generation” paper on inequality of opportunity:
- first generation (theory): moral philosophers and welfare economists Rawls (1971), Dworkin (1981), Arneson (1989) and Cohen (1989), Roemer (1998);
- second generation (measurement): Lefranc et al. (2009), Checchi and Peragine (2010), Bourguignon et al. (2007), Ferreira and Gignoux (2011);
- third generation (econometric specification): Li Donni, Rodríguez, Rosa Dias (2015), Brunori, Hufe and Mahler (2018).

Roemer's Model

$$y_i = g(C_i, e_i)$$

- y_i: individual's i outcome;
- C_i: circumstances beyond individual control;
- e_i: effort.

Types and effort tranches

- Romerian type: set of individuals sharing exactly the same circumstances;

- effort tranche: set of individuals exerting the same effort;

- no random component:
  same type and same tranche → same outcome;

- there is equality of opportunity if and only if:

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

$\Rightarrow$ IOP = within-tranche inequality.

Equality of opportunity: weaker definition

- Van de Gaer (1993): IOP = between-type inequality;
- Van de gaer's approach is the most popular in empirical analysis;
- between-type approach: no need to measure effort.
- measures obtained with the two approaches differ conceptually and empirically.

Effort identification

- effort: observable and not observable choices;
- Roemer's identification strategy, two assumptions:
  1. monotonicity: $\frac{\partial g}{\partial e} \geq 0$
  2. orthogonality: $e \perp\!\!\!\perp C$
- degree of effort = quantile of the type-specific outcome distribution;

3-step estimation

1. identification of Romerian types;
2. measurement of degree of effort exerted;
3. (Roemer's) IOP = within-tranche inequality

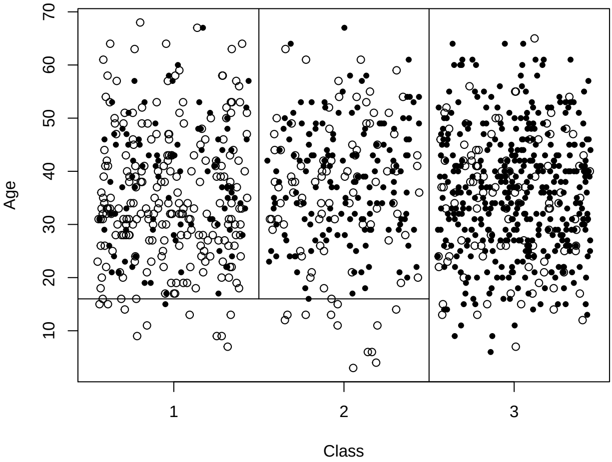
Roemerian types

- first generation papers tried a direct implementation of Roemer's theory;
- unobservable circumstances (downward bias);
- sparsely populated types (upward bias);
- the trade-off is now solved identifying types by estimating a conditional inference regression tree (Hothorn et al, 2006).

Romerian types, cnt

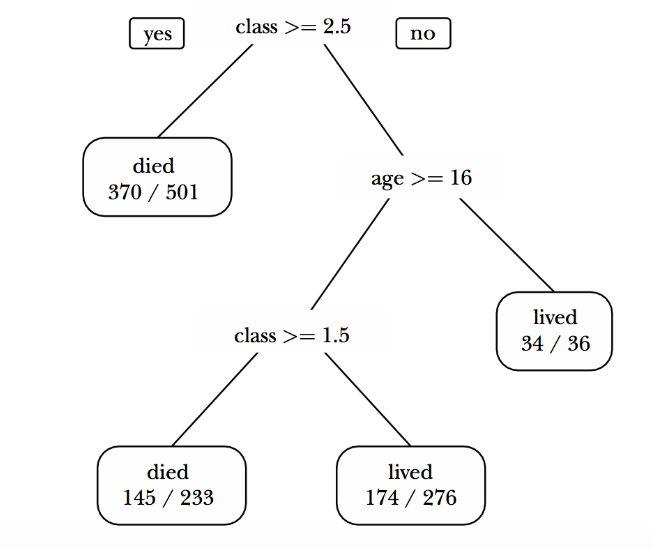
- we use regression tree to identify types;
- a tree predicts a dependent variable based on observable predictors (Morgan and Sonquist,1963)
- the population is divided into non-overlapping subgroups
- prediction of each observation is the the mean value of the dependent variable in the group

What is a tree? cnt.



Source: Varian, 2014

What is a tree? cnt.



Source: Varian, 2014

What is a tree? cnt.

- overfitted models explain perfectly in-sample
- but perform poorly out-of-sample
- different solutions lead to different type of trees

Conditional inference trees

- we use *conditional inference trees* (Hothorn et al., 2006);
- splitting are based on a sequence of statistical test;
- Brunori, Hufe, Mahler (2018): highly interpretable and outperform standard methods to identify types.

The algorithm

- choose α
- $\forall p$ test the null hypothesis of independence:
  $H^{C_p} = D(Y|Cp) = D(Y), \forall C_p \in \mathbf{C}$
- if 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 subsample

Effort

- recall: IOP quantifies to what extent individuals exerting the same degree of effort obtain the same outcome;
- standard approach: choose an arbitrary number of quantiles;
- low efficiency and limited comparability across studies.

Bernstein polynomials

- approximate the ECDF with a polynomial;
- for any quantile $\pi \in [0, 1]$ we can predict the expected outcome in all types;
- we use Bernstein polynomials.

Bernstein polynomials

- introduced in 1912 by Sergei Bernstein
- today: mathematical basis for curves' approximation in computer graphics
- outperform competitors (kernel estimators) in approximating distribution functions (Leblanc, 2012)

Bernstein polynomial of degree 4

$$B_4(x) = \sum_{v=0}^{4} \beta_v b_{v,4}$$

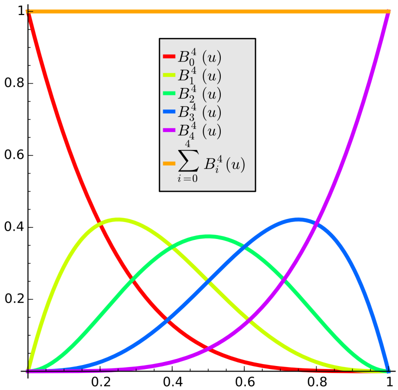
where $b_{v,4}$ is the v-th Bernstein basis polynomial restricted to the interval [0, 1]:

$$b_{v,k} = \binom{k}{v} x^v (1-x)^{k-v}$$

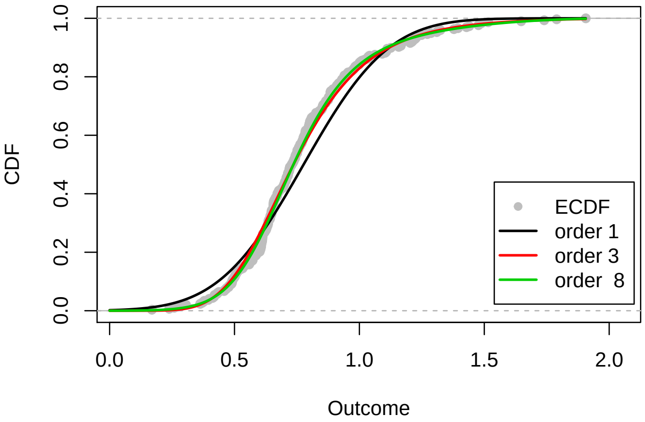
example

$$\begin{aligned} b_{0,4} &= (1-x)^4 \\ b_{1,4} &= 4x(1-x)^3 \\ b_{2,4} &= 6x^2(1-x)^2 \\ b_{3,4} &= 4x^3(1-x) \\ b_{4,4} &= x^4 \end{aligned}$$

Bernstein polynomials, cnt



ECDF approximation by Bernstein polynomials



Choice of the polynomial's degree

- the polynomial is estimated with the *mlt* algorithm written by Hothorn (2018);

- out-of-sample log-likelihood to select the most appropriate order of the polynomial;

IOP estimation

- Knowing the shape of all type-specific distribution functions we can estimate the distribution of 'unfair' inequality

- $IOP = Gini\left(\frac{y_i}{\mu_j}\right)$, μ_j expected outcome at percentile j;

- no longer need to choose a particular number of effort quantiles;

- order of polynomial varies across types to maximize estimate reliability.

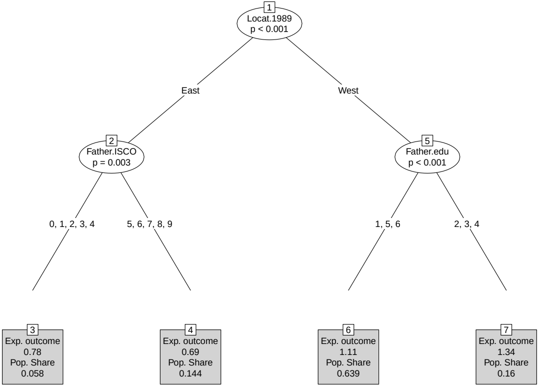
Data

- SOEP (v33) including all subsamples apart from the refugee samples;
- adult individuals (30-60);
- y = age-adjusted household equivalent disposable income;

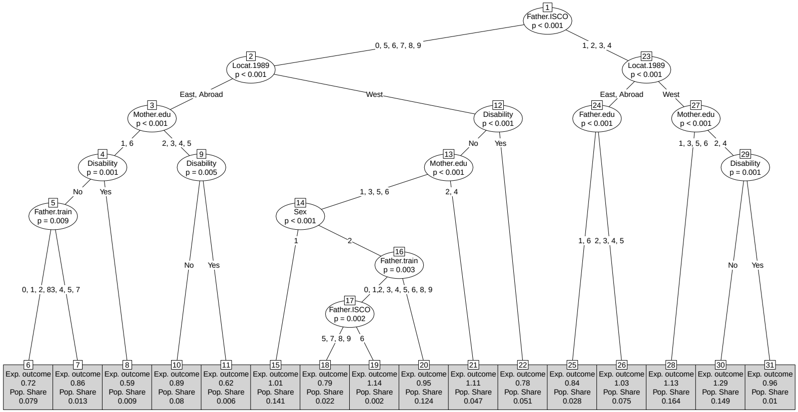
Data, cnt.

- SOEP provides comprehensive information about circumstances beyond individual control;
- waves considered 1992-2016;
- circumstances considered: migration background, location in 1989, mother's education, father's education, father's occupation, father's training, month of birth, disability, siblings;

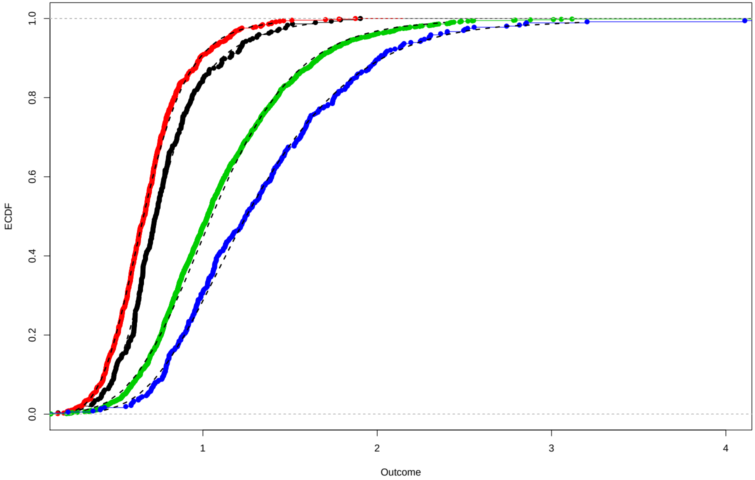
Opportunity tree in 1992



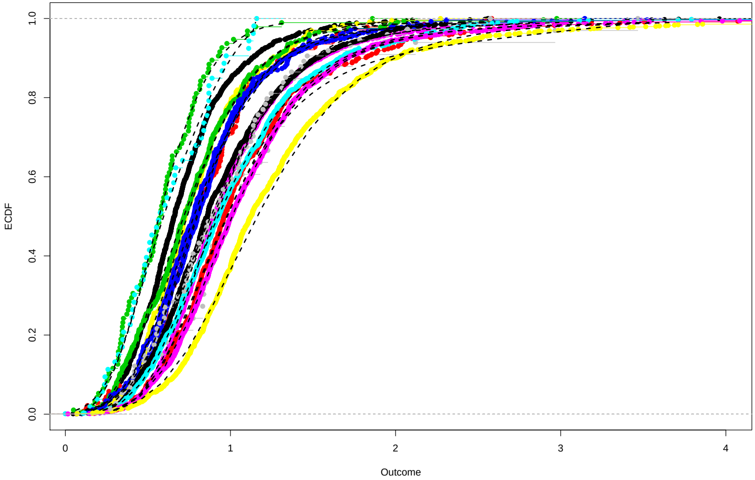
Opportunity tree in 2016



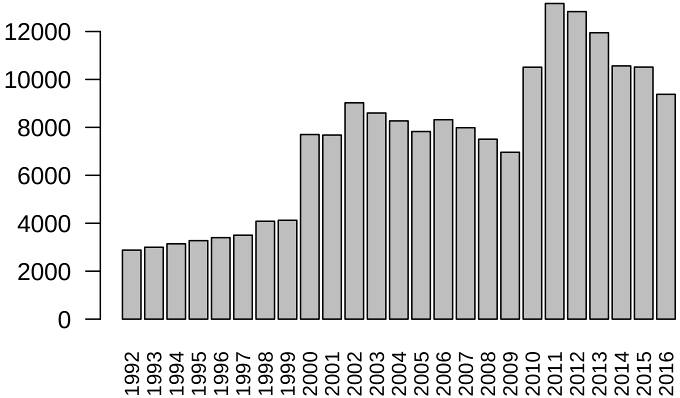
IOP in 1992



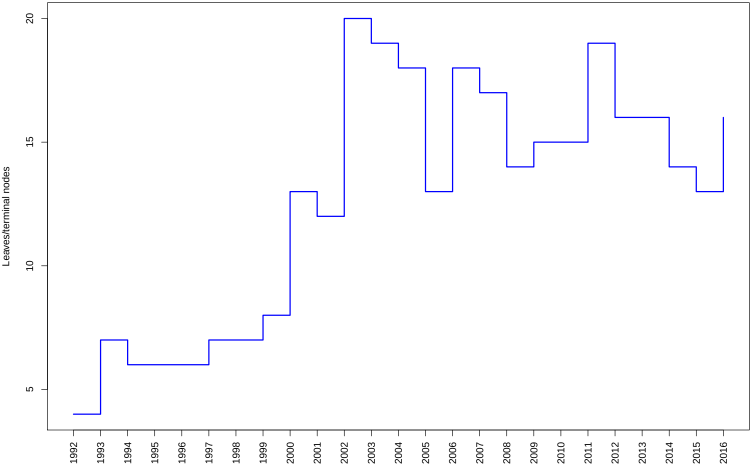
IOP in 2016



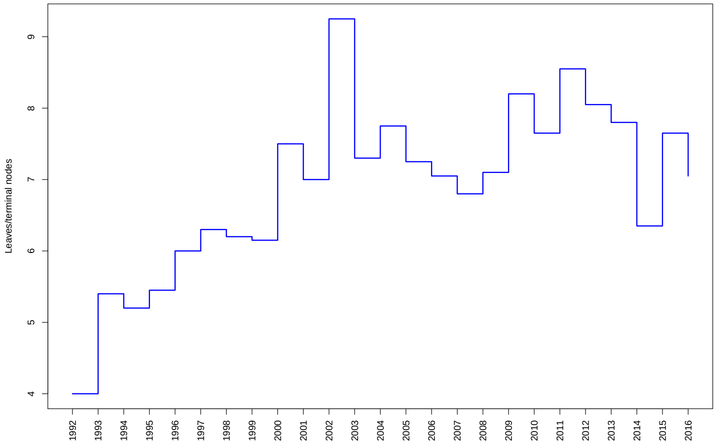
Sample size 1992-2016



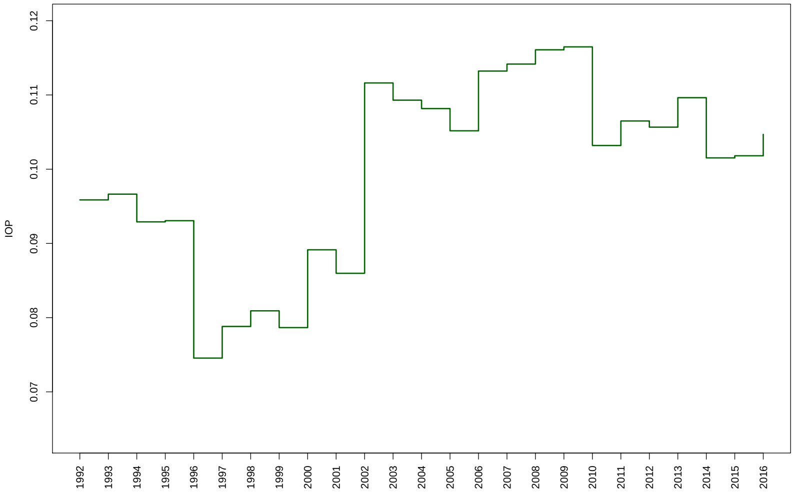
Number of types 1992-2016



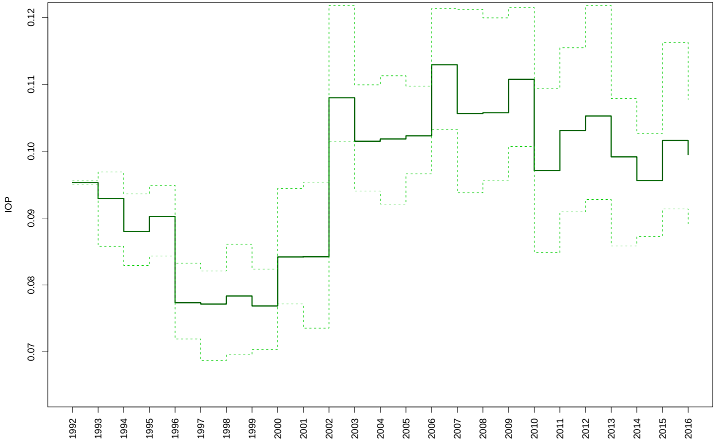
Mean number of types (same sample size) 1992-2016



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.

Summary

- we propose an approach to estimate IOP fully consistent to Roemer's theory;
- effort identification method maximizes efficiency and comparability;
- since 1992 in Germany the opportunity structure has become more complex;
- IOP declined after reunification and increased with Hartz reforms;
- is today about 10% higher than in 1992.