GEOM Database

Paolo Brunori[a,b] and Pedro Salas-Rojo[a]

[a]III London School of Economics, [b]University of Florence

May, 2023

Description of the Database

Measure IOp in "as many countries and years as possible".

- Provide a global understanding of the IOp phenomenon.
- Measure how circumstances contribute.
- Grasp time trends.
- Provide a standpoint framework and methods.

Steps

- Develop methods and scripts (R).
- Design code architecture (R).
- Data collection and cleaning (R + Stata).
- Production of estimates and checks (R).
- Web design (Python).

Output

- IOp estimates.
- Circumstance importance.
- Plots.
- Technical documentation.
- Country-specific data reports.
- Package to get estimates and plots.

Main Settings

- Outcome: equivalized household income (2010 USD).
- Circumstances: sex, ethnicity, parental occupation, parental education, place of birth.
- Individuals between 18 and 80 years.
- Control by age of household head.

Europe: minor issues

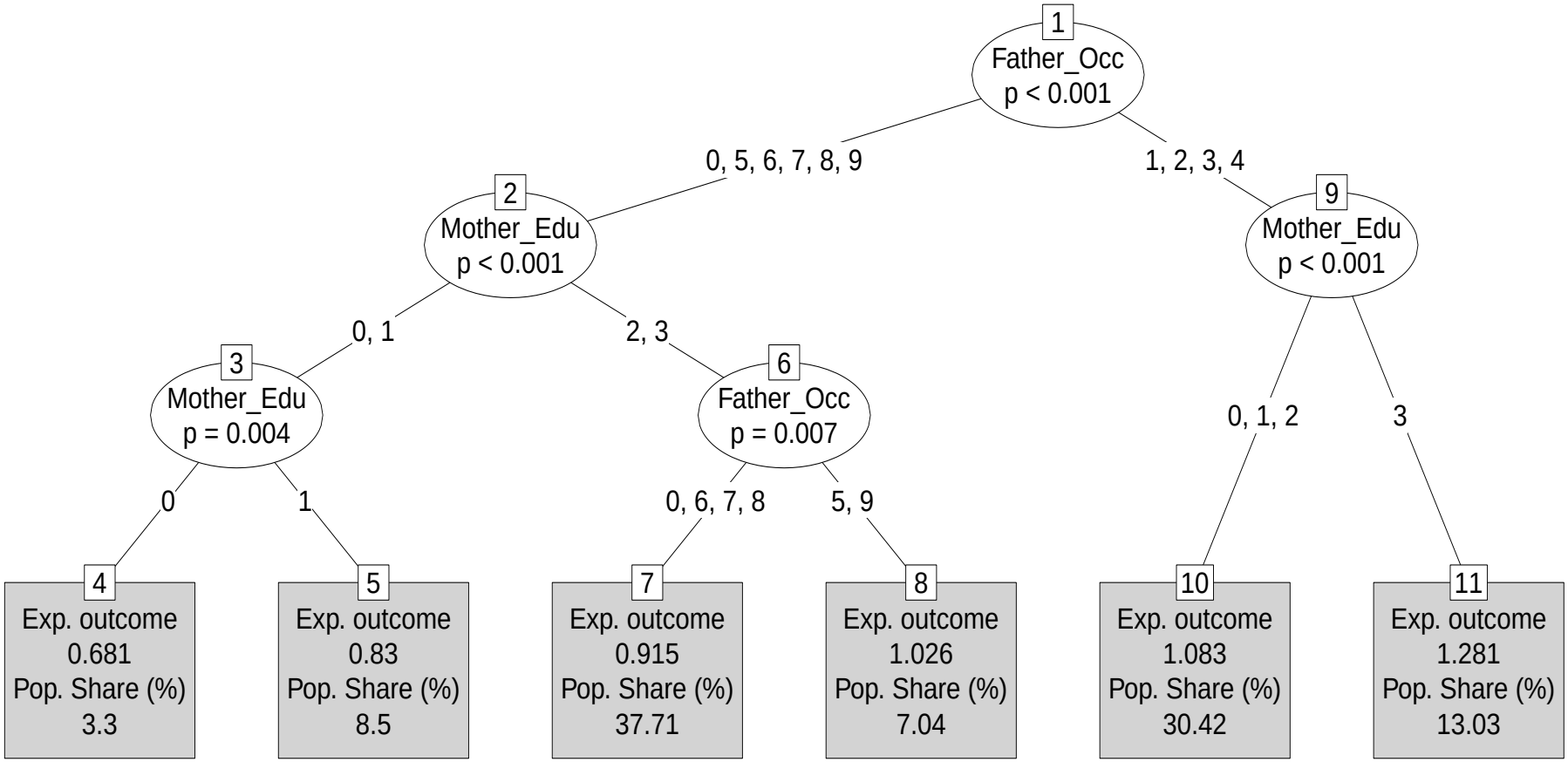
EUSILC (2005, 2011, 2019).

- No ethnicity.
- Place of birth poorly recorded in some waves.
- Some countries have small sample size.

IOp in Italy (2019): 0.11 Gini points, around 34% of total inequality.

Ex-ante Tree in Germany (2019)

IOp: 0.09 Gini points, around 31% of total inequality.



LATAM: minor issues

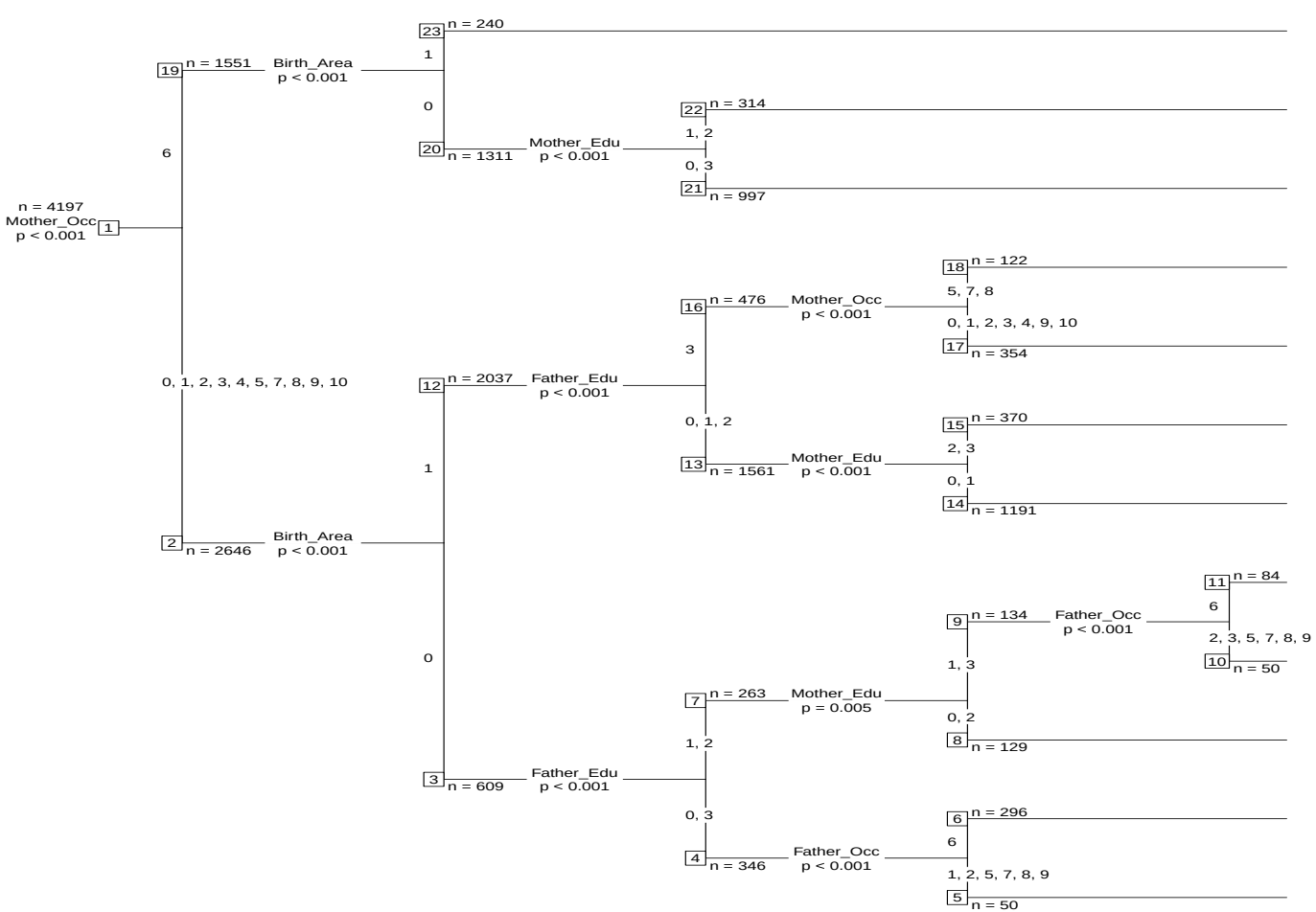
Data from Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Peru, Panama and Guatemala.

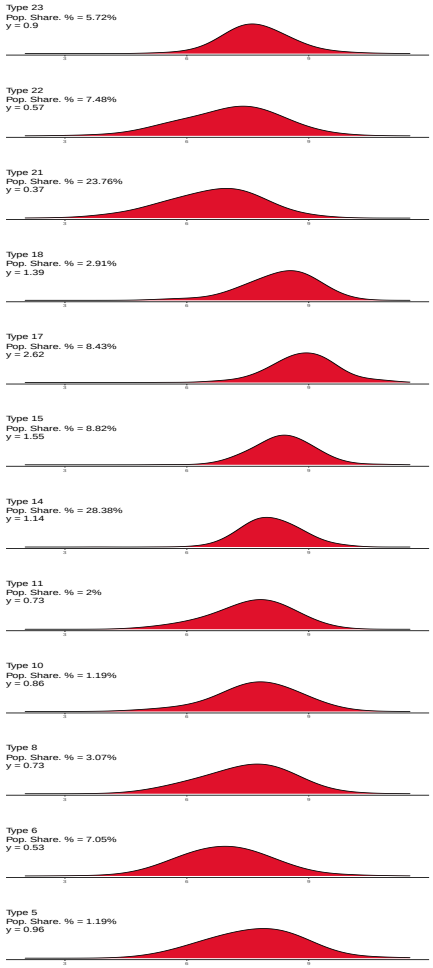
- Different sample size.
- Some circumstances are not available.
- Different codings in categories.
- Place of birth sometimes within countries.

IOp in Brazil (2014): 0.32 Gini points, around 66% of total inequality.

Ex-post Tree in Bolivia (2008)

IOp: 0.23 Gini points, around 46% of total inequality.





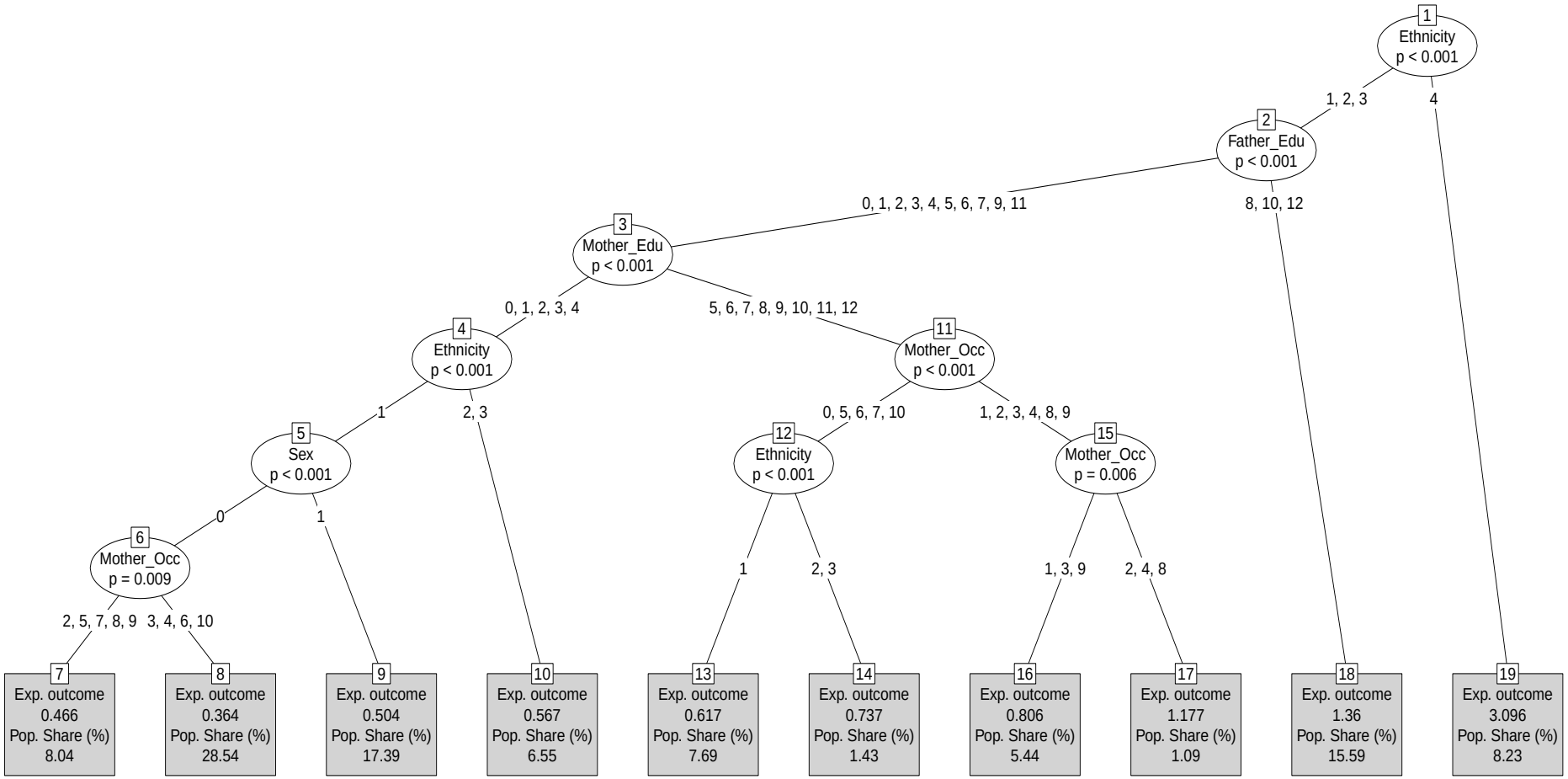
Africa: many problems

Data from South Africa, Uganda, Ghana, Tanzania, Sierra Leona, Mali, Tongo, Niger, Senegal, Benin, Burkina Faso, ...

- Data scarcity: few countries collect circumstances.
- No incomes: use consumption.
- Sample sizes may be small, we are flexible with age constraints.

IOp in South Africa (2017): 0.42 Gini points, around 71% of total inequality.

Ex-ante Tree in South Africa (2017)



Asia: limited data availability

Data from China, India, Indonesia, Bangladesh, Pakistan, ...

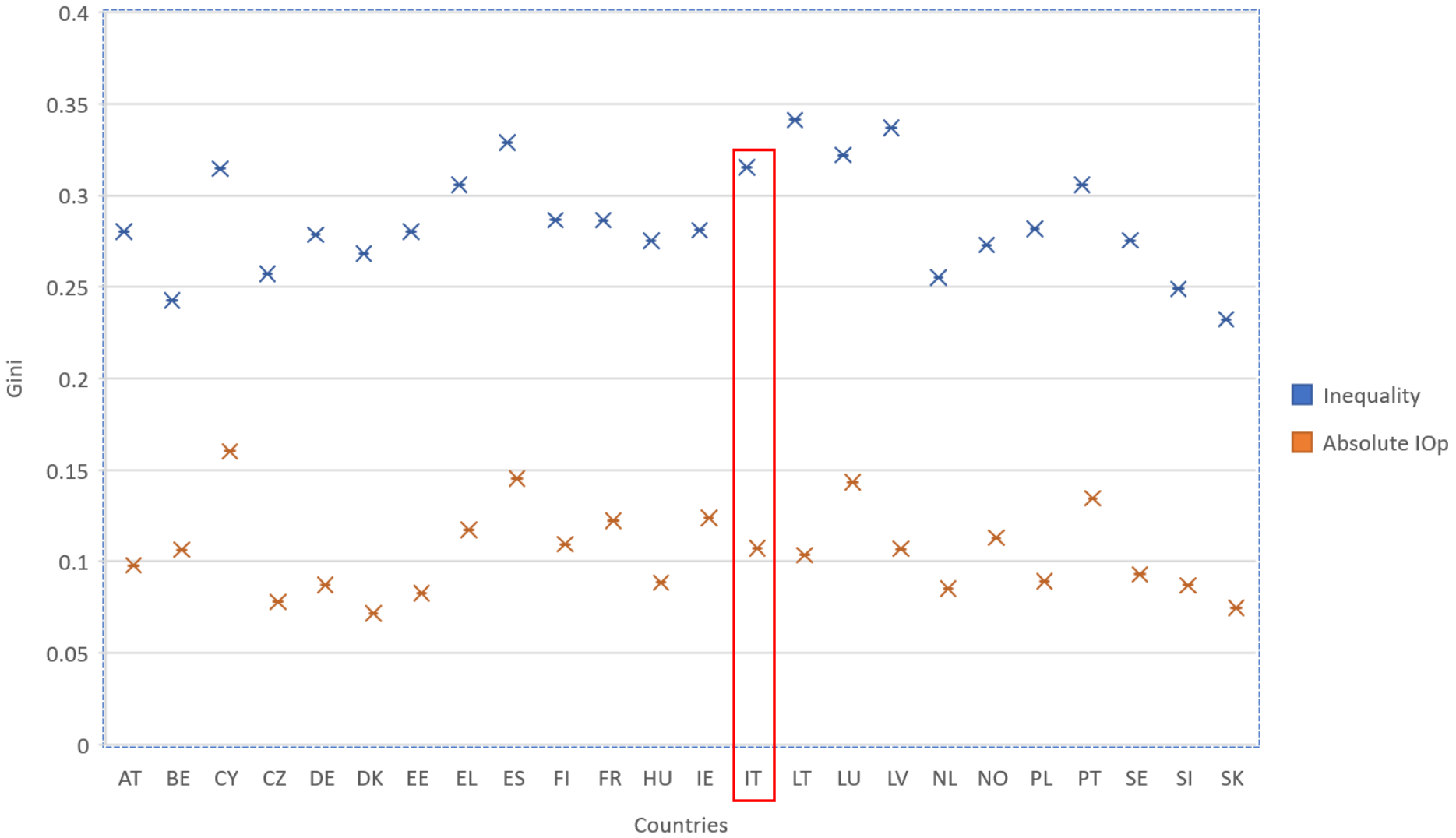
- Complex process of homogenization.
- Surveys are representative of regions or areas.
- Good news: for China we have a panel!

No estimates yet to be shown!

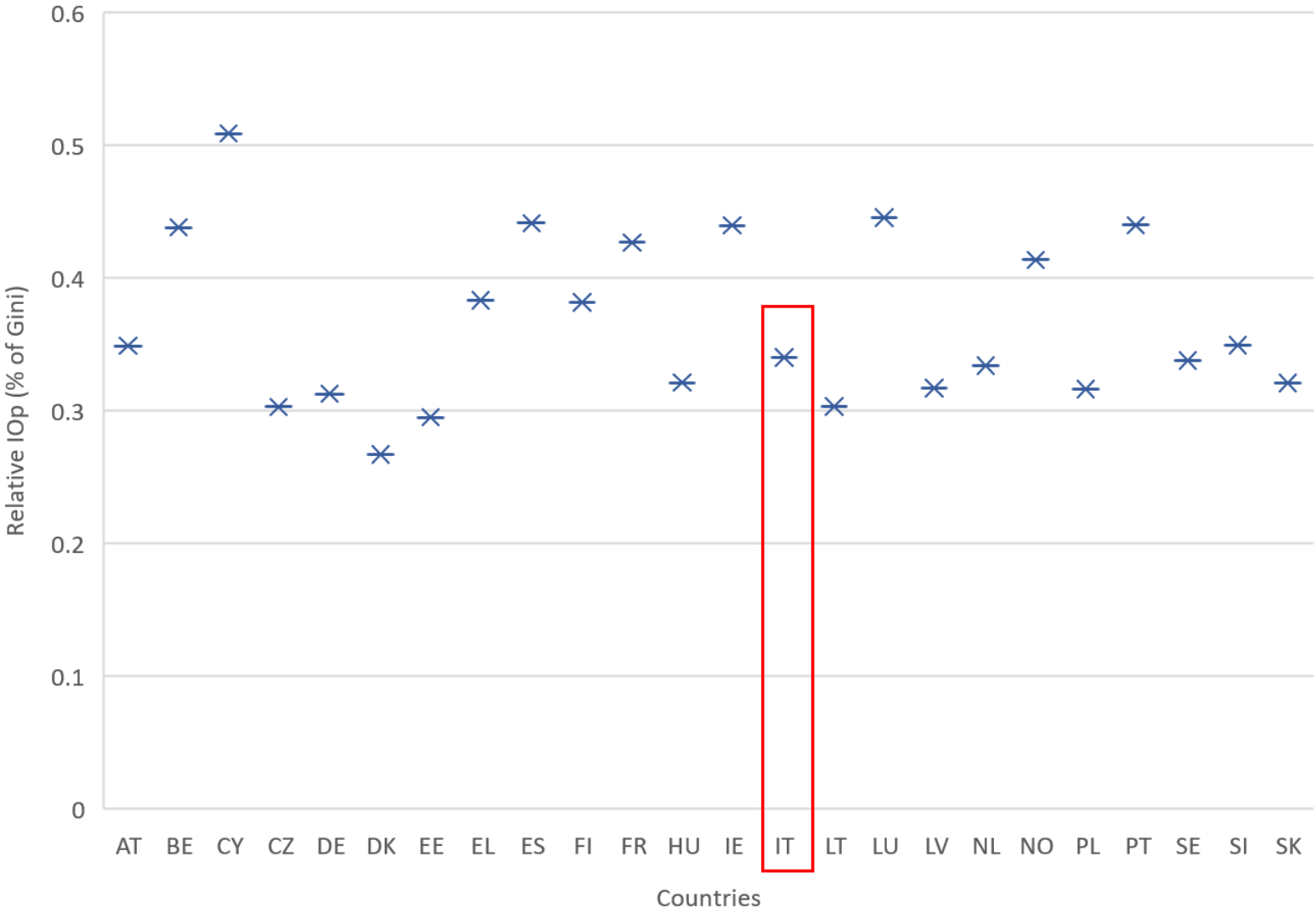
Other countries and regions

- Some are quite complete: United States, Australia and Korea.
- Others are so far not available: Japan, Turkey, New Zealand, Russia, Caribbean, Arab countries, Middle Asia...

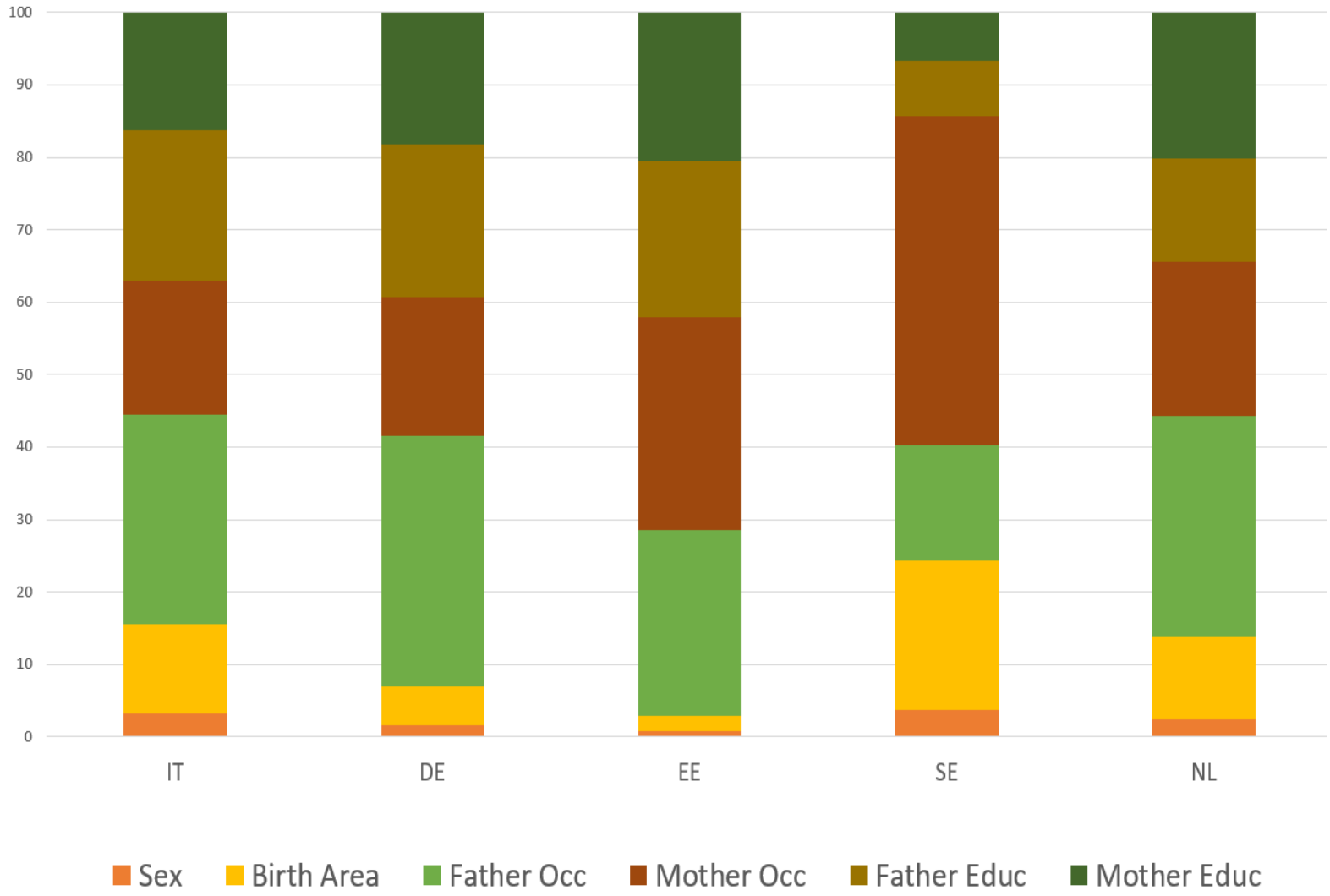
IOp estimates in Europe (2019)



Relative IOp in Europe (2019)



Circumstance Importance (2019)



Problem Set (STATA)

We challenge you to estimate (ex ante) IOp with real data!

- Play around with data.
- Launch trees and random forests.
- Estimate IOp.
- Explore differences between both methods

The Data

European Social Survey (2018) in Italy.

Outcome: education, health.

Circumstances: education and occupation of parents, sex, country of birth, parental country of birth.

Other: age, region of residence

More information is provided in the codebook.

The Methods

Classification and Regression Trees (CART) and Random Forests.

- Use the "crtrees" package in Stata (documentation included)
- More information in https://ideas.repec.org/c/boc/bocode/s458573.html

Task I: Estimate IOp with Regression Tree

Try to:

- Estimate trees with outcomes and circumstances.
- Measure IOp with several sets of circumstances.
- Explore how the algorithm is tuned.
- Play with parameters.

Is IOp sensitive to the circumstances and parameters employed? Why? How can this affect policy implications or public debate?

Task II: Estimate IOp with Random Forests

- Estimate Random Forests with outcomes and circumstances.
- Measure IOp with several sets of circumstances.
- Compare with Tree results
- Explore how the algorithm is tuned
- Play with other parameters.

Is IOp sensitive to the circumstances and parameters employed? How can this affect policy implications or public debate?

Example CART

Estimate a tree, predicting health with isco_f and educomp_f

Overall Inequality: 0.7797 (sd health)

Absolute IOp: 0.2062 (sd y_hat)

Relative IOp: 26.44%

Example Random Forest

Estimate a random forest, predicting health with isco_f and educomp_f

Overall Inequality: 0.7797 (sd health)

Absolute IOp: 0.1550 (sd y_hat)

Relative IOp: 19.88%

Thanks a lot!

Happy to discuss or help: p.salas-rojo@lse.ac.uk