Roots and Evolution of Unfair Educational Inequality in Spain:
A Normative Approach with Machine Learning

Carlos Gil

UNIVERSITÀ DEGLI STUDI FIRENZE

Pedro Salas-Rojo

LSE International Inequalities Institute

FES FEDERACIÓN ESPAÑOLA DE SOCIOLOGÍA

CI06 COMITÉ DE INVESTIGACIÓN DESIGUALDAD Y ESTRATIFICACIÓN SOCIAL

Madrid, April 8-9 2025

X Reunión Intercongresos CI-06

Limitations: Inequality of [Educational] Opportunity (IoP)

1. Unsystematic normative conceptualization (Martínez García and Giovine 2025; Grätz 2024).

2. Multiple estimation approaches (Strömberg and Engzell 2023; Marqués-Perales et al. 2023).

3. Focus on educational attainment vs performance (Fernández-Mellizo 2022; 2014).

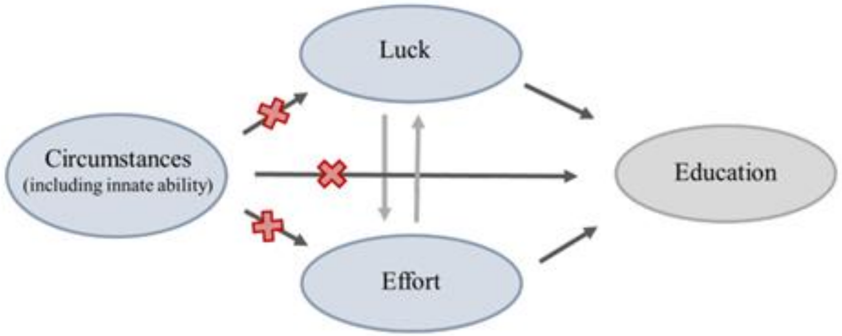
Background: IoP Measurement Approaches

(Björklund and Jäntti, 2020)

- Relative intergenerational mobility
  - Log-linear (Breen and Müller 2020), risk ratio, rank-rank correlation (income), surnames (Clark 2014)
  - Single origin variable (class, education, income)

- Sibling models (total family effect)
  - Sibling correlation in education (Grätz et al. 2021) ≈ [0.4 - 0.5] → all family-constant circumstances, but black box
  - Group-specific contribution (between-group) to overall distribution/correlation (Karlson and In 2024)
  - Still, (increasing) limited sibling data given low fertility rates in young cohorts

- Twin models (total family effect net of genetics)
  - ACE variance decomposition model (Baier et al. 2022) → C = black box
  - External validity issues



✓ (Unfair) Inequality of opportunity (ascribed factors' types)

- *Luck egalitarianism* (Roemer and Trannoy 2016):
  - Unfair: types of ascribed circumstances beyond individual's control (genes, sex, ethnicity, parental class)
  - Fair: effort ⫫ circumstances
- OLS, latent class analysis, regression trees, random forests (Brunori, Ferreira and Salas-Rojo 2023)

Aim & Contributions

1. **Normative Formalization:** Informed analysis of unfair inequalities in academic performance from Roemer's (1998) *luck egalitarianism* theory → Theory into practice

2. **Machine Learning Approach:** Transformation trees (Hothorn and Zeileis 2021). Data-driven identification of complex ascribed types (intersectionality) → Less estimation bias

3. **Feature Importance:** Mapping the relative importance of 8 ascribed circumstances over time → Mechanisms

4. **Trends:** Big dataset comprising 7 waves (2003-2022) of the PISA study → 20 years of IoP

Data & Variables

- PISA (OECD) 7 waves (2003, 2006, 2009, 2012, 2015, 2018, 2022)

- Total analytical sample = 152,446 students (wave mean ≈ 22,000, sd ≈ 8,000)

- Outcome (*y*): test scores in math competence domain at age 15-16 (mean ≈ 500, sd ≈ 100)
  - Plausible values (5-10)

- 8 Socio-demographic (unfair) Circumstances (*c*):
  - Sex (2 [1. Female; 2. Male])
  - Mother/father education (5 [0. ISCED 0, 1. ISCED 1, 2. ISCED 2, 3. ISCED 3ABC, ISCED 4, 4. ISCED 5, 5A, 5B, 6])
  - Mother/father occupation (11 [10-ISCO 1-digit + inactive])
  - Immigration status (3 [1. Native, 2. Second-generation, 3. First-generation])
  - Language at home (2 [1. Test language; 2. Other])
  - School community size (5 [1. A village or rural area < 3,000 - 5. A large city > 1,000,000])

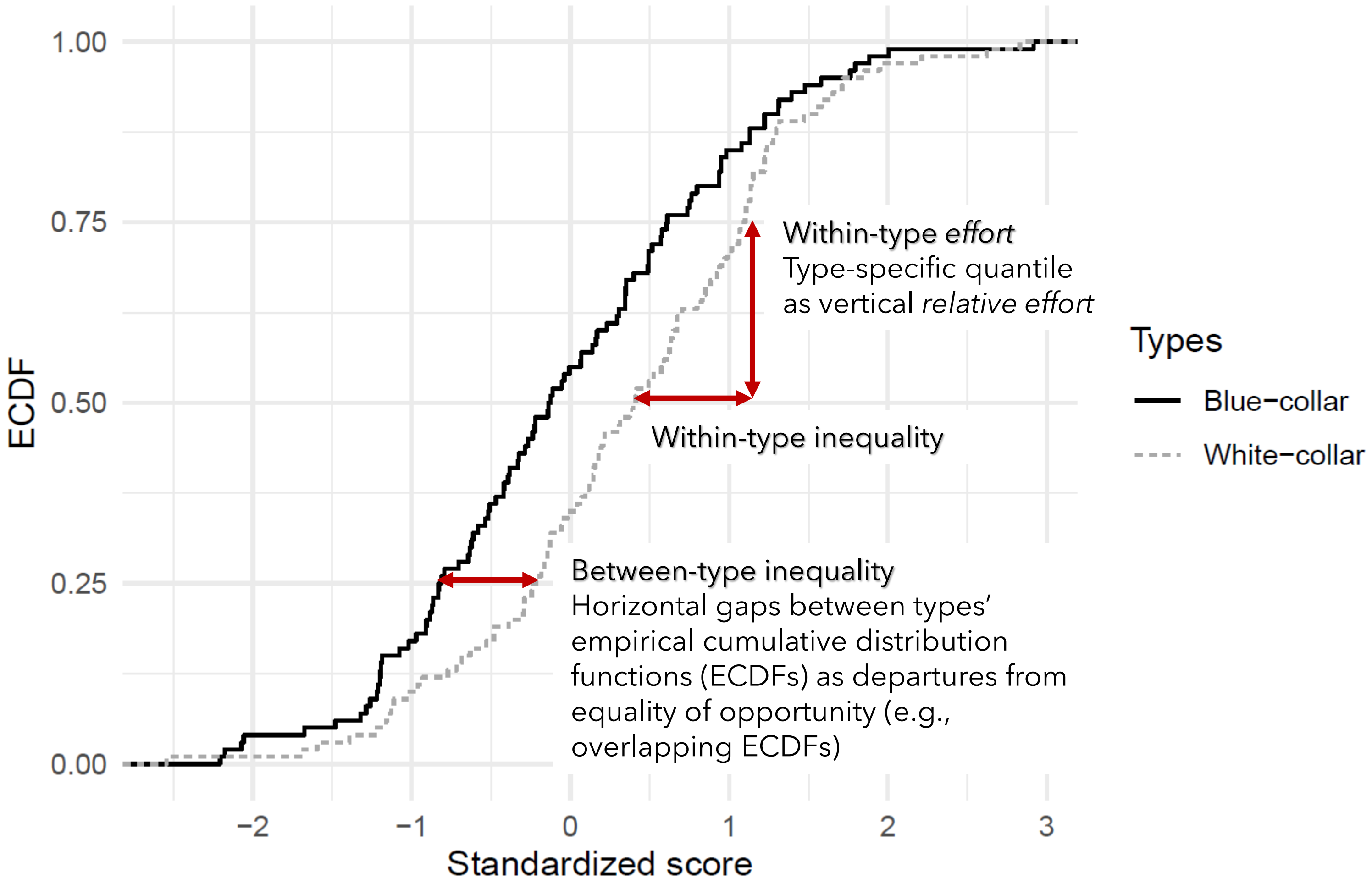
Methods: IoP Definition

> "Thus, when comparing the efforts of individuals of different types, we should somehow adjust for the fact that those efforts come from distributions that are different–a difference for which individuals should not be held responsible"
> Roemer (1998:458)

$$I\left(\hat{y}_{cf}\right) = \int_{q=0}^{1} I_q\left(y_q - \mu_q\right)dq$$

$$IOP = \frac{I\left(\hat{y}_{cf}\right)}{I(y)}$$

Simplified Population → 1 Circumstance & 2 Types



Within-type *effort*
Type-specific quantile as vertical *relative effort*

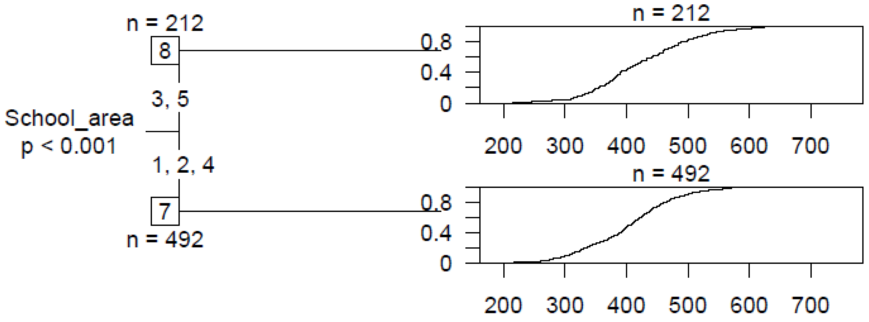
Within-type inequality

Between-type inequality
Horizontal gaps between types' empirical cumulative distribution functions (ECDFs) as departures from equality of opportunity (e.g., overlapping ECDFs)

Methods: Identifying Types

- Optimal data-driven model specification vs. arbitrary decisions (Brunori et al. 2023):
  - Lower-bound bias: Data availability and omitted variables **(few large types)**; mean outcome (random forests) → - variance / + bias
  - Upper-bound bias: Overfitting by multiple variables and interactions with small n **(many small types)** → - bias / + variance

- Trading-off biases with ML-transformation trees (Hothorn and Zeileis 2021):
  - Ex-post approach: outcome distribution beyond mean inequality
  - Recursive optimal sample partitions (C splits/interactions) fitting the data: largest gap in outcome ECDFs between 2 subgroups of C
  - Trees' high variance → Forest bootstrapping random *k* samples (n/50)



Machine Learning vs OLS: Bias-Variance Trade-off

R^2 including 3-way interactions of the 8 circumstances (2003-2022 average)

Model	**R^2 In-Sample** (80%)	**R^2 Out-of-Sample** (20%)
OLS	0.26	0.03
Random Forest	0.26	0.16

Data: PISA-Spain

- **OLS overfits:** high in-sample R^2 (low bias), but nearly zero predictive power on new data (high variance).
- **Random Forest generalizes better:** same in-sample R^2, but much higher out-of-sample R^2 (low variance).

Types: Mean Transformation Tree (2022)

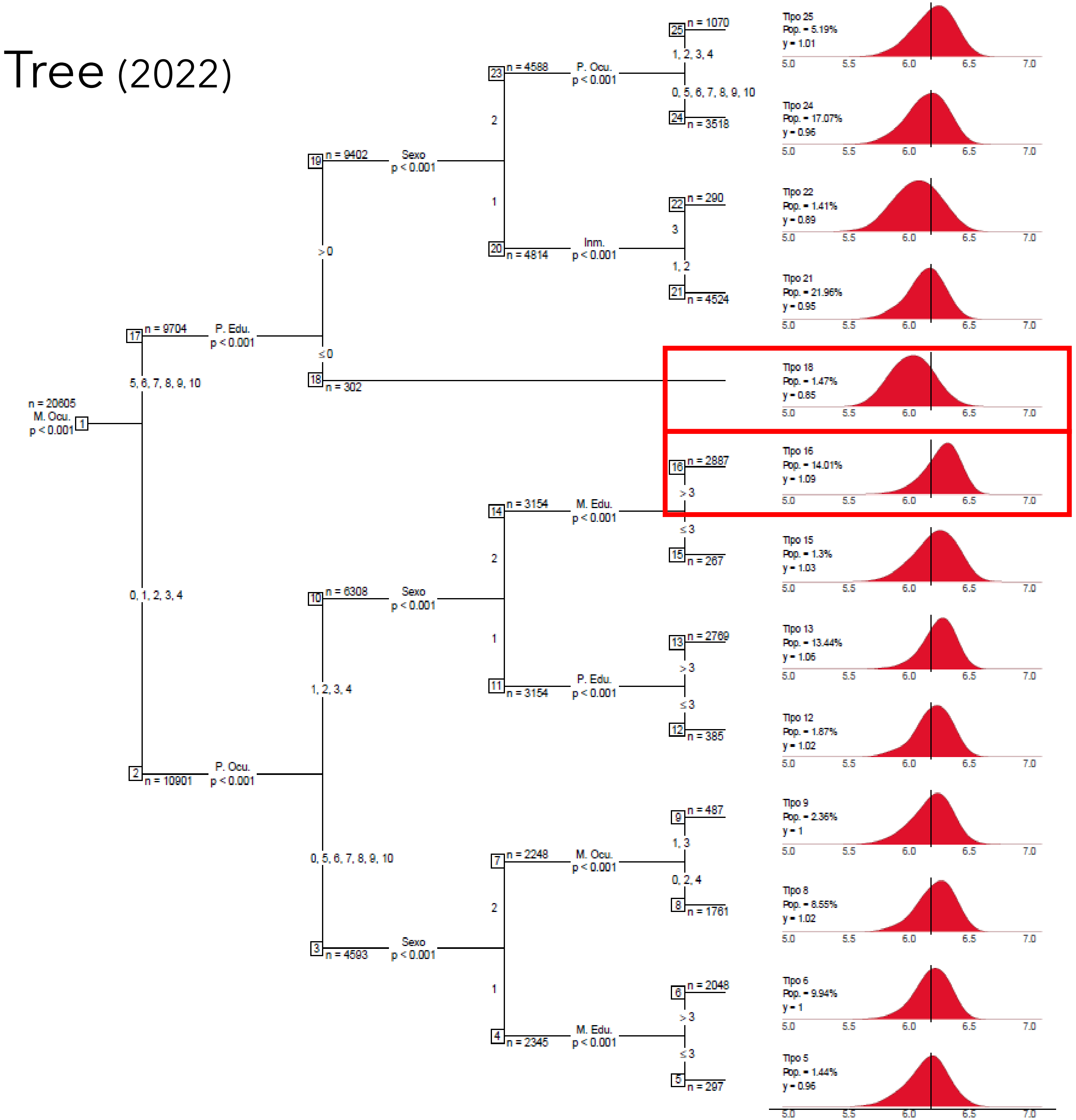
N types = 13
R² = IoP Variance / Total Variance = 12 %

Worst-off Type (#18)
1.5% of sample
y = 85% below mean

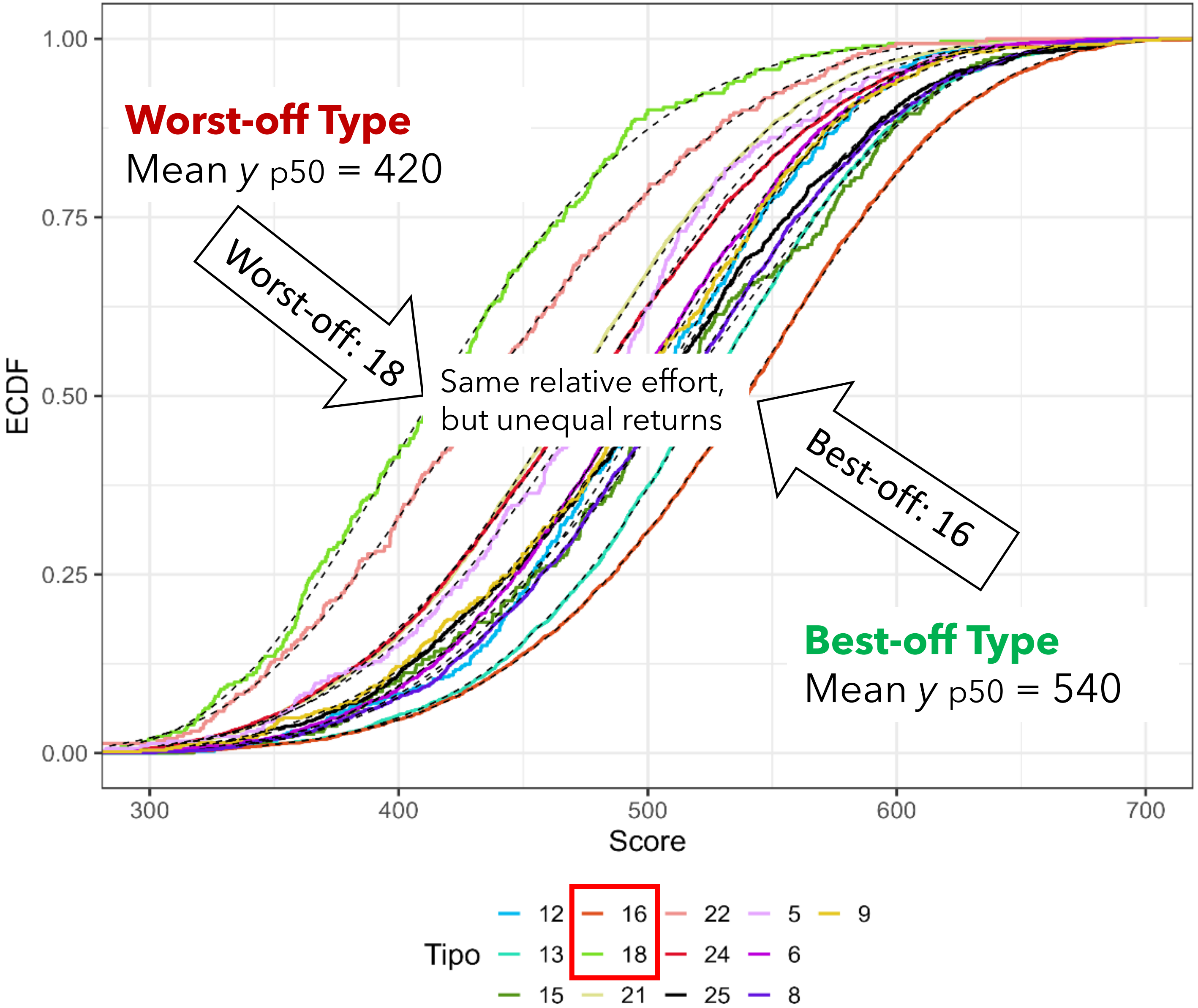
Mother Occupation: ISCO 5-10 →
Father Education: ISCED 0

Best-off Type (#16)
14% of sample
y = 9% above mean

Mother Occupation: ISCO 1-4 →
Father Occupation: ISCO 1-4 →
Sex: Men →
Mother Education: ISCED > 3



Types: Empirical Cumulative Distribution Functions (2022)



Worst-off Type
Mean *y* p50 = 420

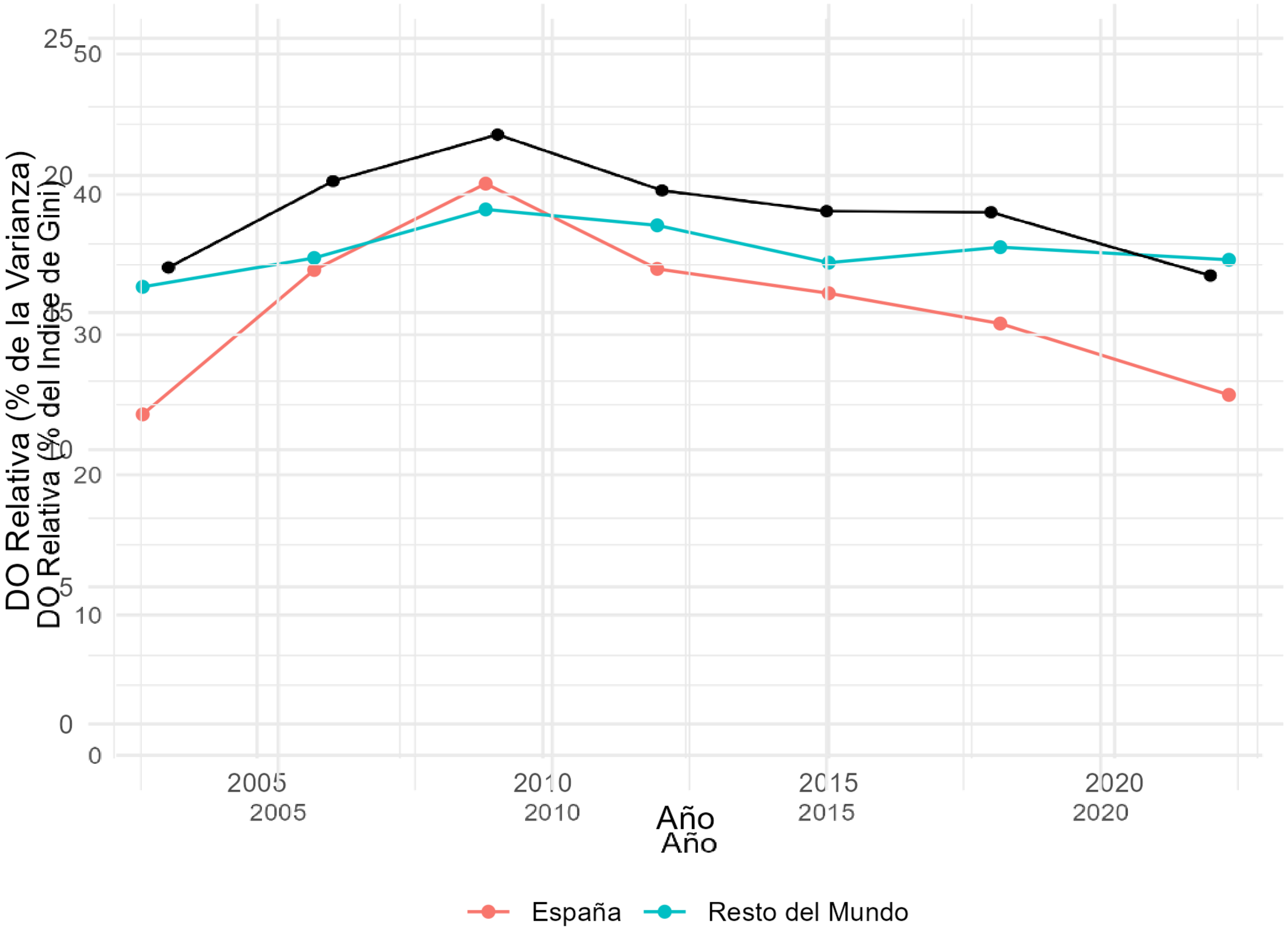
Worst-off: 18

Same relative effort, but unequal returns

Best-off: 16

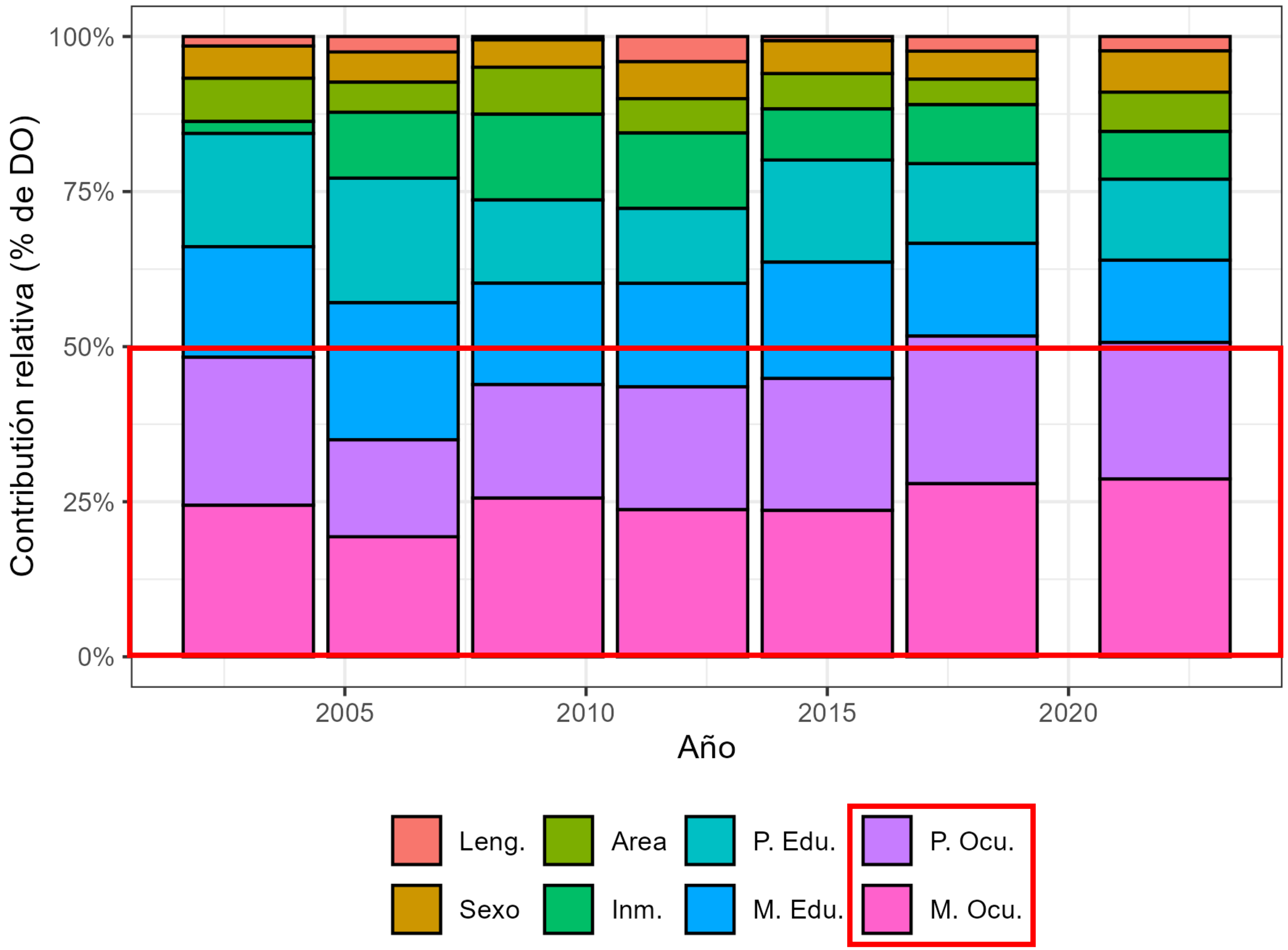
Best-off Type
Mean *y* p50 = 540

IoP Trends: % Variance & Gini



Shapley-Shorrocks Decomposition: Feature Importance over Time

✓ Bootstrapped standardized contribution of a variable c to predicted inequality (reduction) when c is omitted from the prediction (tree), averaged across all possible combinations of circumstances that omit c (Shorrocks, 2013).



Conclusion

✓ Robust (theory-and data-driven) formalization of unfair IoP, ranking circumstances:

- Identifying complex intersectional types without overfitting vs OLS
- Sensible alternative lacking sibling data
- Flexible approach: adding circumstances if available (parental wealth/income, genetics…)

✓ Constant (inverted U-shaped) IoP over time, consistent with persistent attainment inequality in Spain

✓ Persistence in the key circumstances explaining IoP (social inequality structure)

- Parental SES (occupation/education) contributes more to IoP than other ascribed factors (migrant origin; sex)
- Declining role of parental education in IoP, in line with expansion and negative selection (Valdés 2022)

Thanks for your attention! ☺

carlos.gil@unifi.it

Limitations

- Tiny predictive improvement (R^2) vs standard (parametric OLS), but less error, variance and arbitrary

- Lower-bound IoP (0.2) vs. sibling models (0.4) → variables (un)available (genes, wealth)

- Effort (within-type *y* quantile: fair ↕ inequality) → Strong model assumptions and normative implications

Next Steps

- ➢ Replicating with reading scores

- ➢ EoP: *Rawlsian maxi-min* (e.g., worst-off types distance)

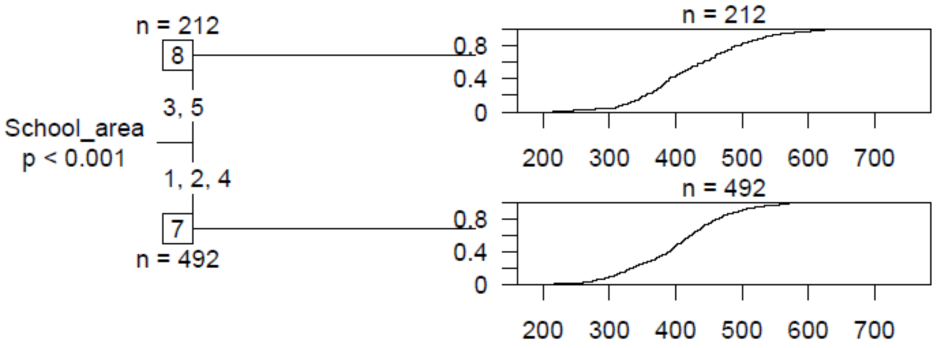
Transformation Trees: Identifying Types

(Brunori, Salas-Rojo and Ferreira 2023; Hothorn and Zeileis 2021)

- Unconditional outcome distribution → Regressors' space (circumstances) optimal partition to maximize the likelihood of fitting the data;

- Binary splitting variables → Maximize shape differences in the conditional distribution functions (CDFs) with a linear polynomial approximation

Fitted conditional outcome distributions across types



7-step Algorithm

1. Set a confidence level (1 – α [0.01]);

2. Set a polynomial order (*m*) reducing the out-of-sample log-likelihood (5-fold cross-validation);

3. Estimate the unconditional distribution with the Bernstein polynomial of order *m*;

4. Test the null hypothesis of polynomial parameters stability over all possible binary partitions based on each χ and store p-values;

5. If in each possible partition the Bonferroni-adjusted p-value > α , stop;

6. Otherwise (p-value < α), choose the variable and the splitting value producing the smallest p-value to obtain two subgroups;

7. Repeat step 4:6 for the resulting subgroups.

Shapley-Shorrocks Decomposition: Steps

A. Standardize to 1 (most important circumstance -> relative ratio/contribution)

B. Draw a subsample of the full sample

C. Estimate IOp in this subsample, but setting α = 1;

D. Further, estimate IOp in the subsample for all possible permutation sequences that eliminate circumstance c. This elimination is performed by replacing c with a constant vector 1

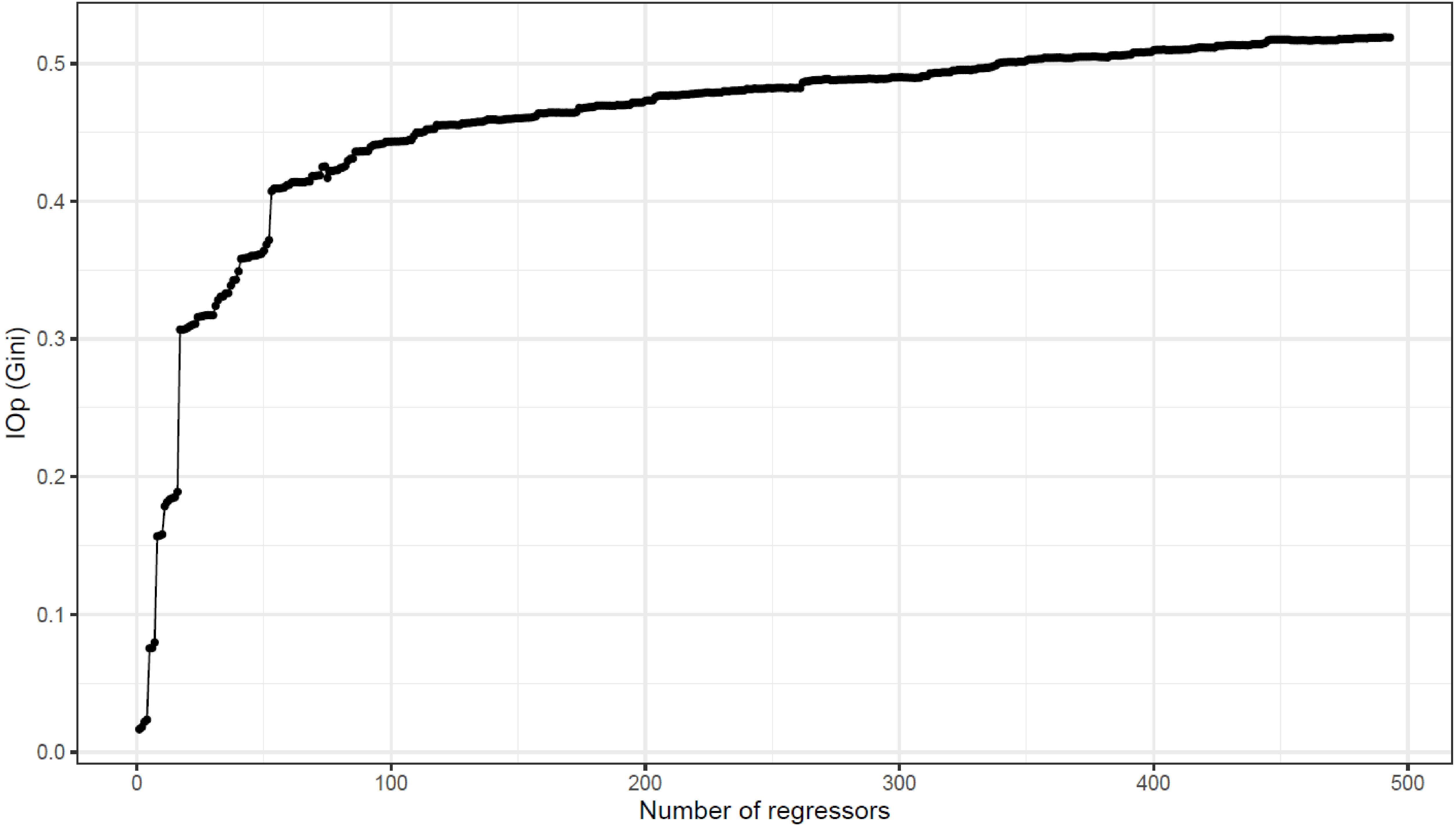
E. Estimate a tree and IOp after each elimination sequence and store results

F. Average IOp across all permutation sequences. The difference between overall IOp and this average is the specific contribution of c;

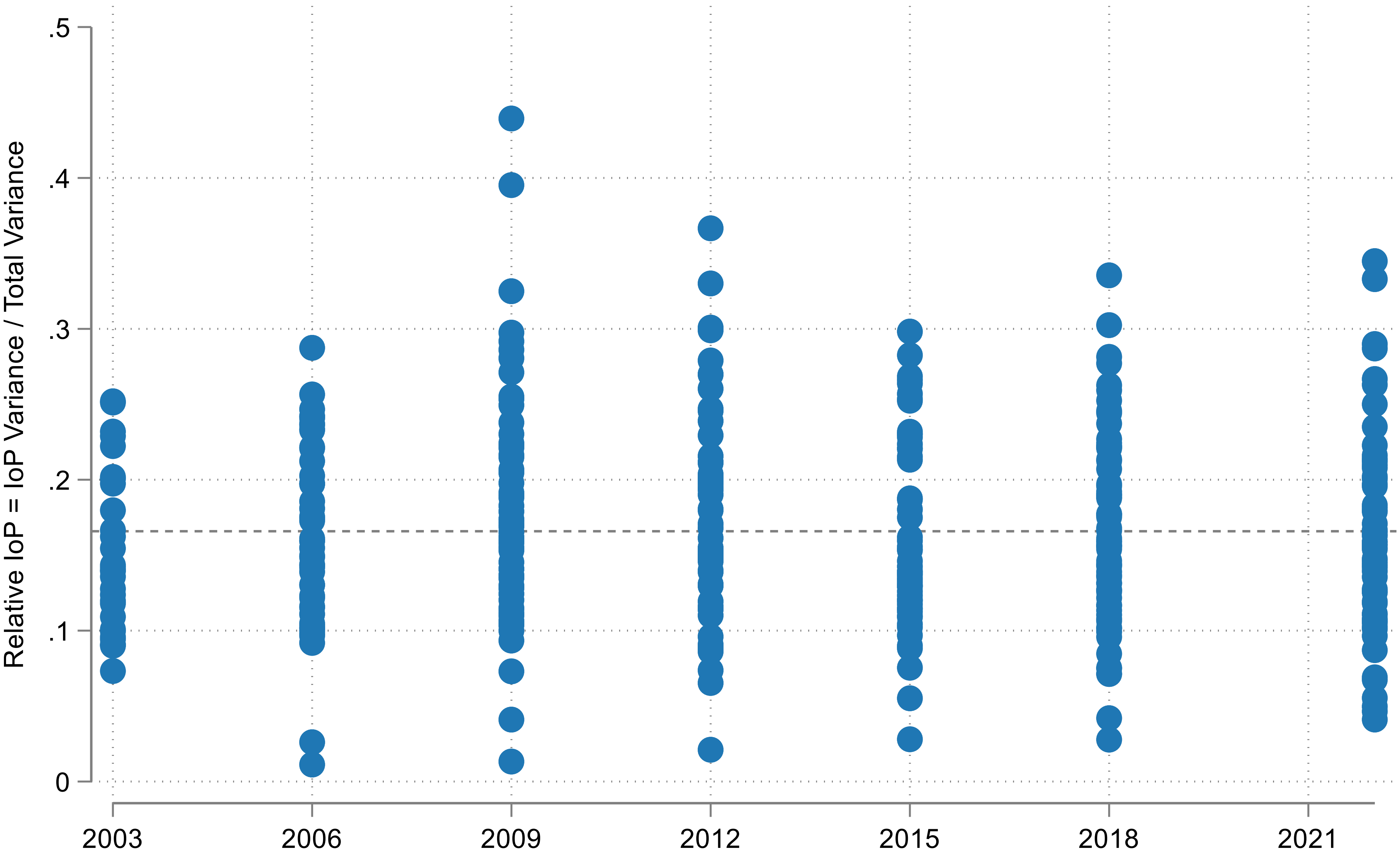
G. Repeat steps A-E z times, to account for different potential data-generating processes. In our case, we set z = 100;

H. Estimate the contribution of c to IOp as the average contribution across these z repetitions;

I. Repeat the algorithm for each c, $k \in \{1, \dots, K\}$.


IOp (Gini)
0.0
0.1
0.2
0.3
0.4
0.5
Number of regressors
0
100
200
300
400
500

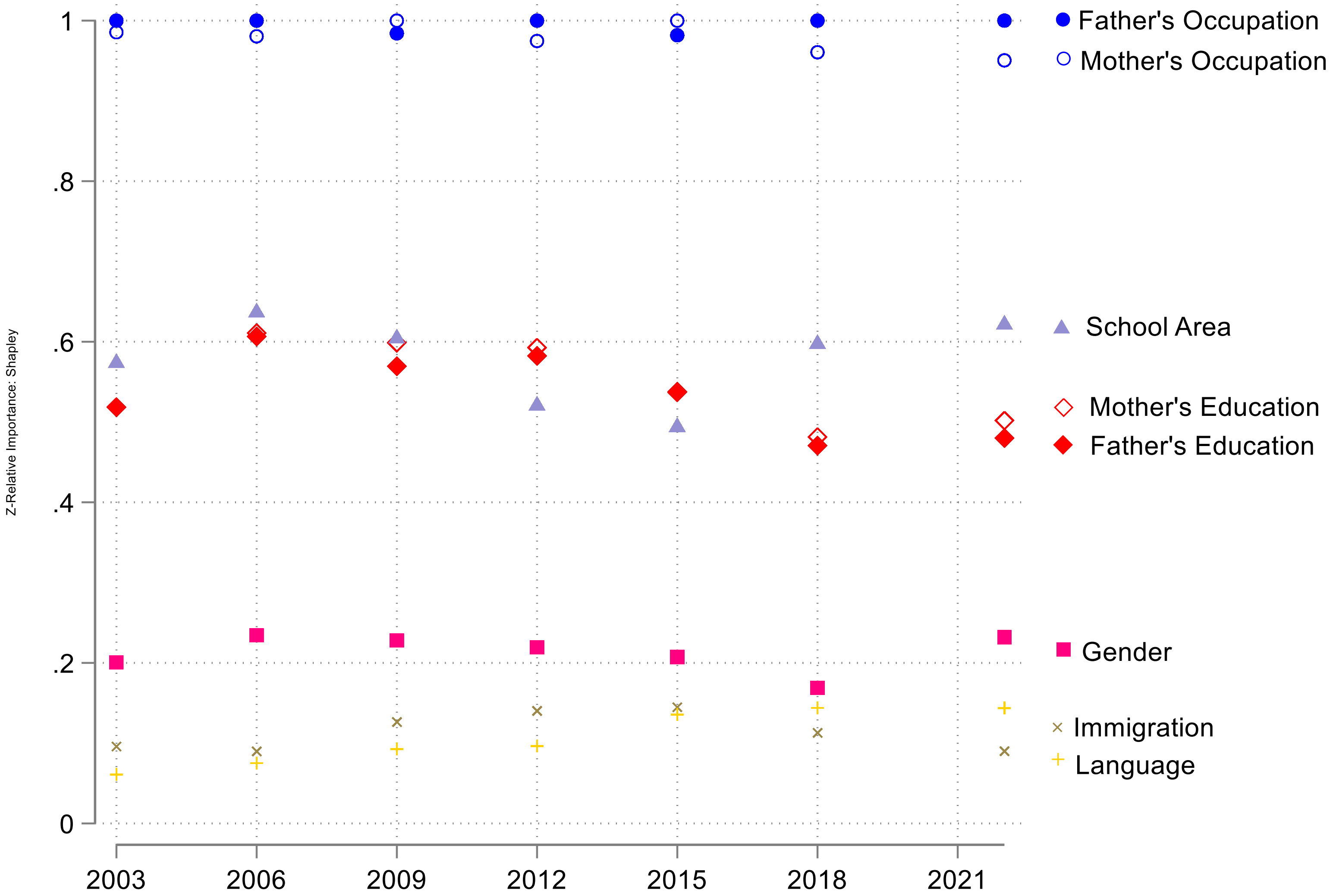
% Unfair IoP Over Time by Countries



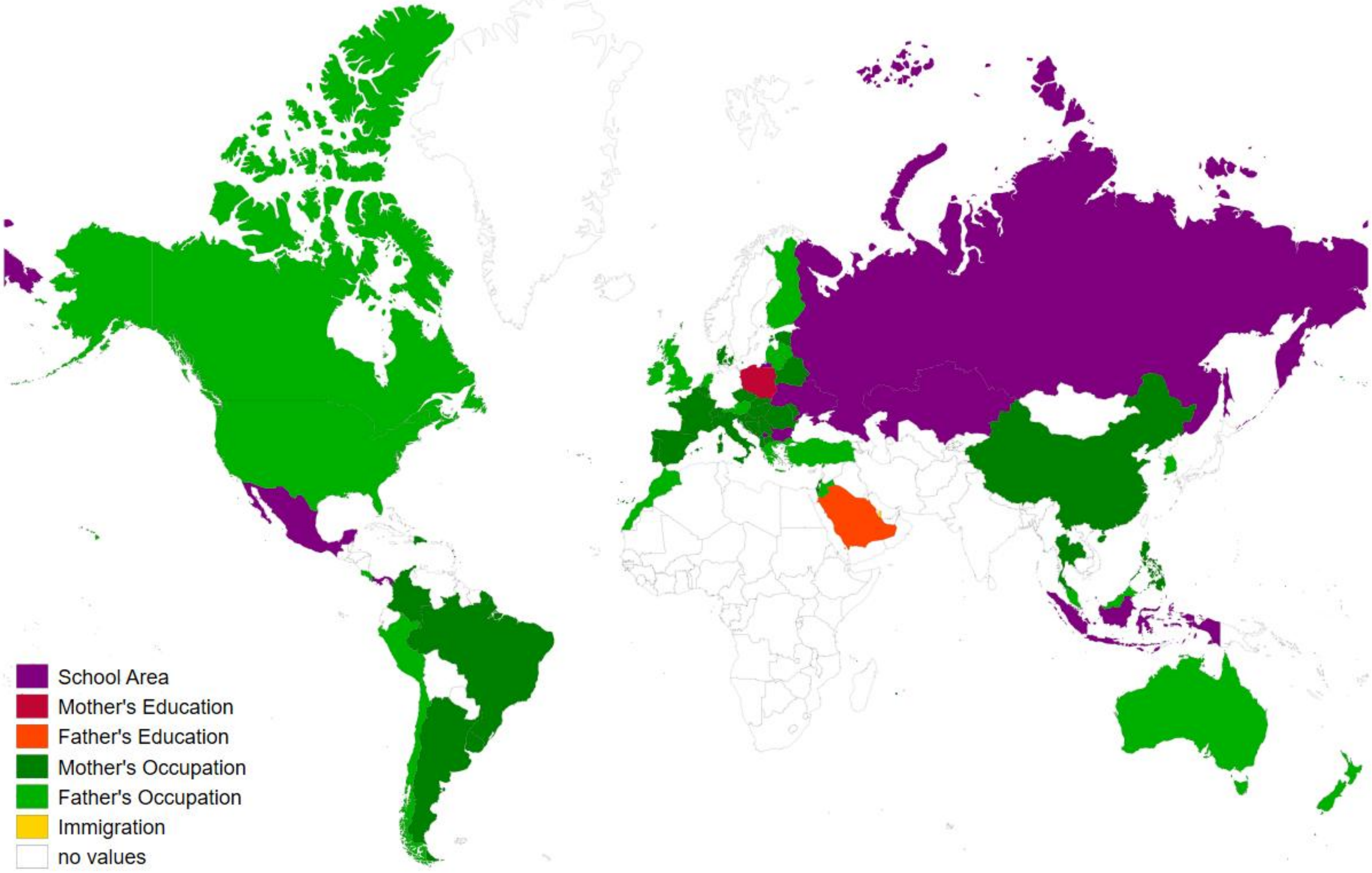
Shapley-Shorrocks Decomposition: Mean Feature Importance Worldwide

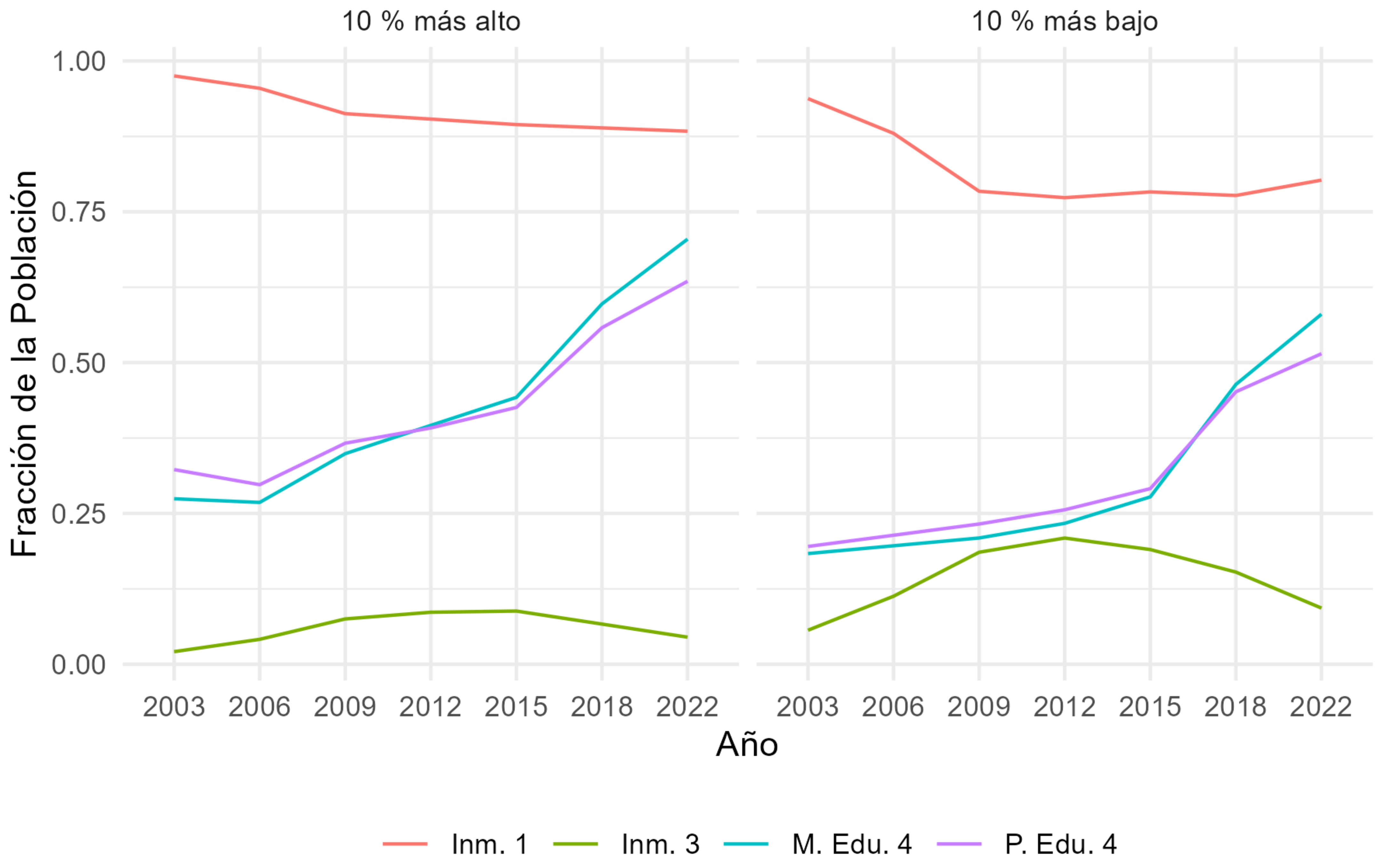
✓ Bootstrapped standardized contribution of a variable c to predicted inequality (reduction) when c is omitted from the prediction (tree), averaged across all possible combinations of circumstances that omit c (Shorrocks, 2013).

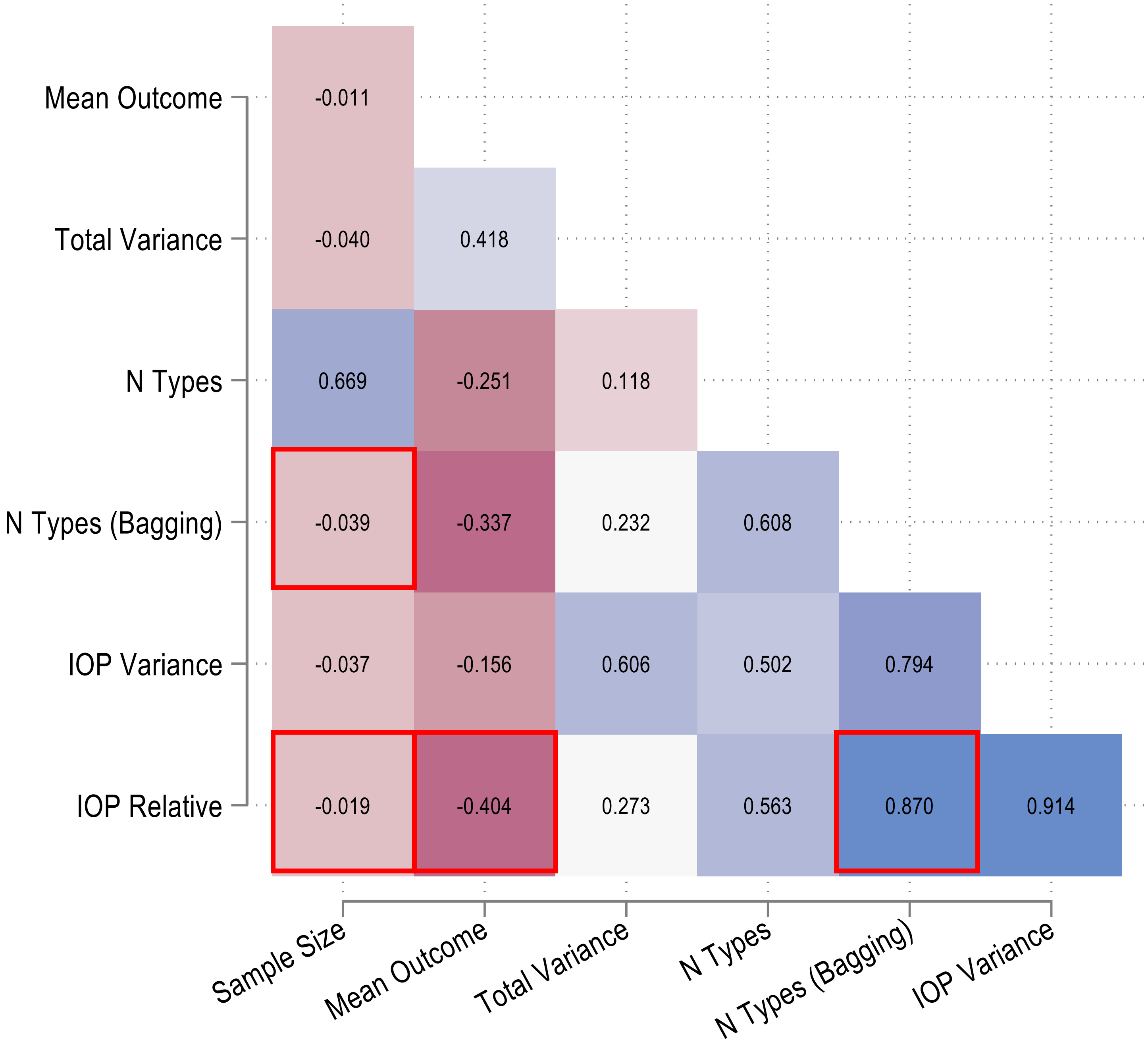
IoP explained variance by *C* vs. the most important circumstance (1)

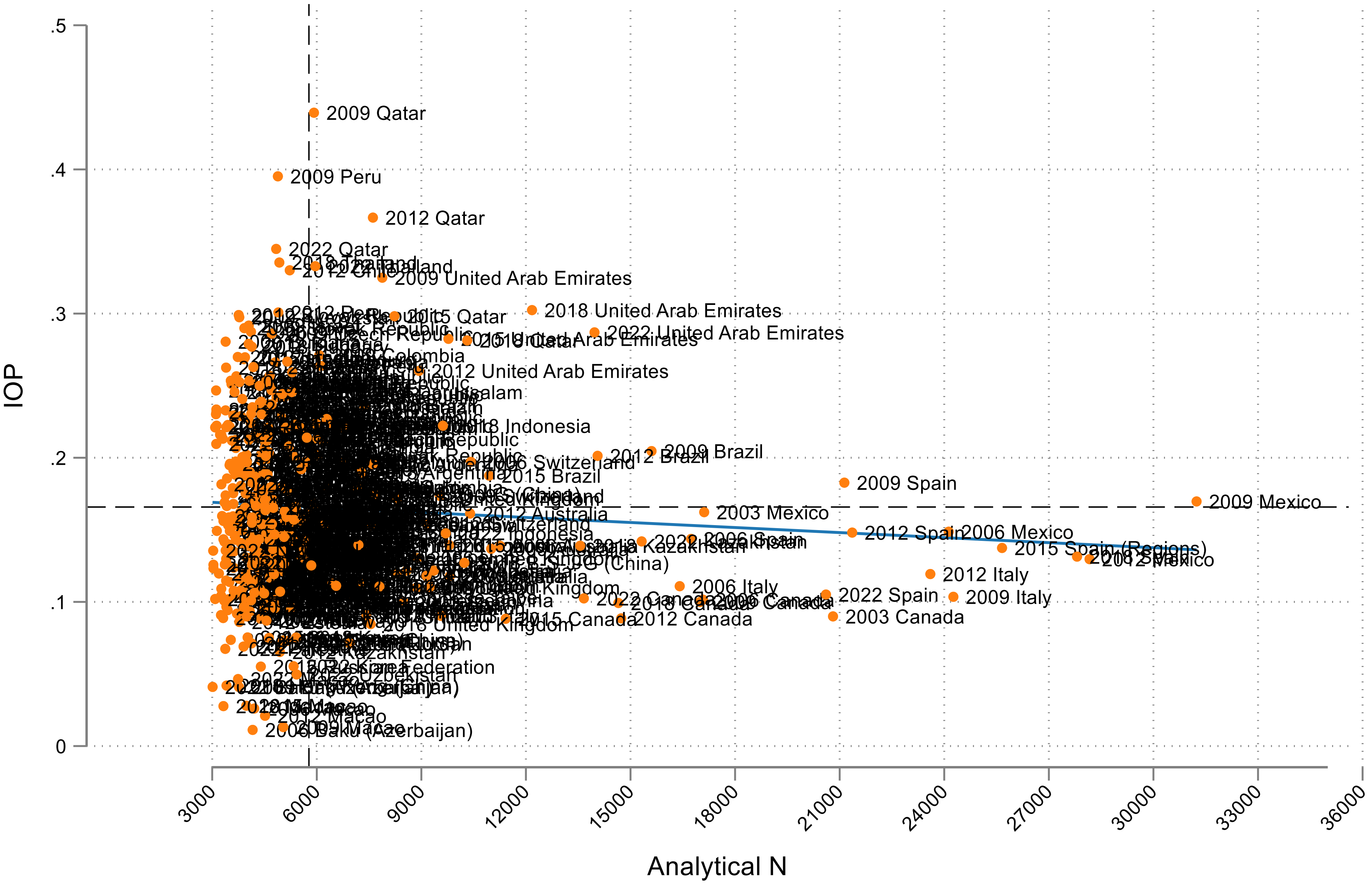


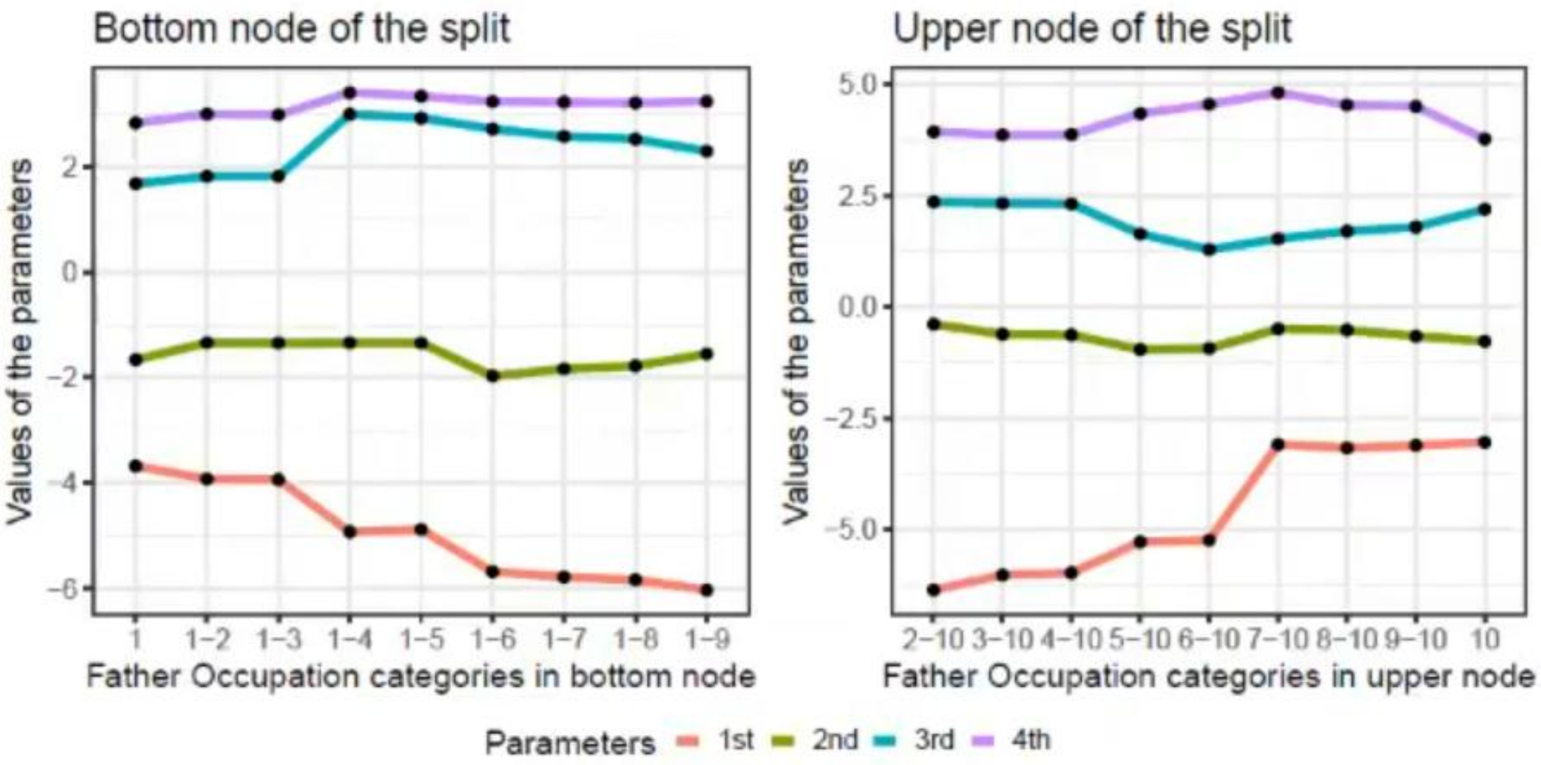
Most Important Circumstance Worldwide (2018)

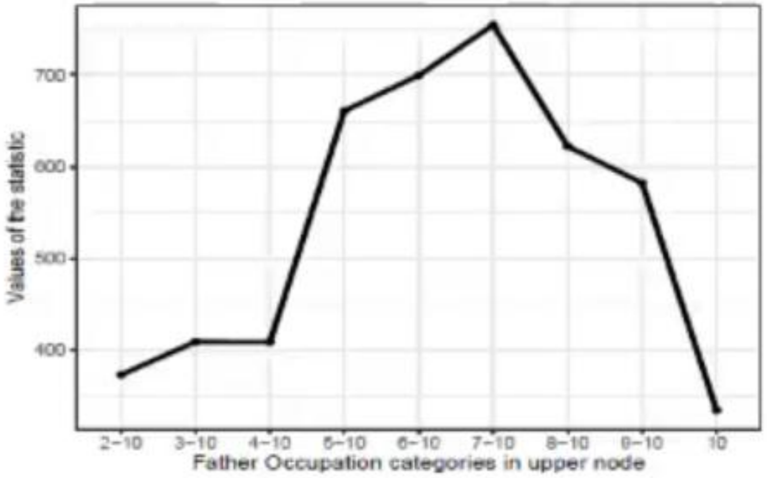



10 % más alto
10 % más bajo
Fracción de la Población
1.00
0.75
0.50
0.25
0.00
2003
2006
2009
2012
2015
2018
2022
Año
Inm. 1
Inm. 3
M. Edu. 4
P. Edu. 4

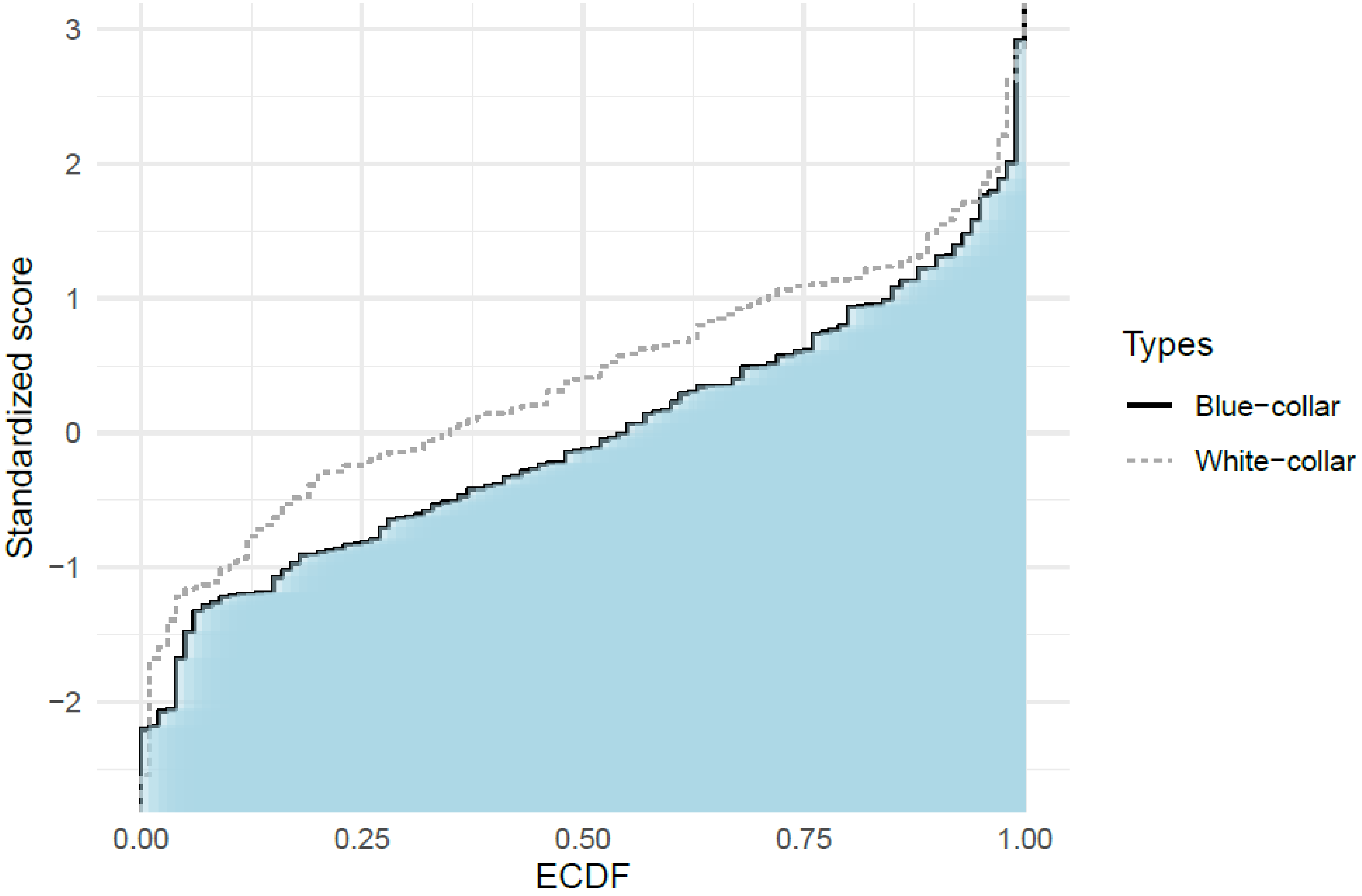

Mean Outcome
-0.011
Total Variance
-0.040
0.418
N Types
0.669
-0.251
0.118
N Types (Bagging)
-0.039
-0.337
0.232
0.608
IOP Variance
-0.037
-0.156
0.606
0.502
0.794
IOP Relative
-0.019
-0.404
0.273
0.563
0.870
0.914
Sample Size
Mean Outcome
Total Variance
N Types
N Types (Bagging)
IOP Variance


IOP
.5
.4
.3
.2
.1
0
Analytical N
3000
6000
9000
12000
15000
18000
21000
24000
27000
30000
33000
36000
2009 Qatar
2009 Peru
2012 Qatar
2022 Qatar
2009 United Arab Emirates
2018 United Arab Emirates
2015 Qatar
2022 United Arab Emirates
2018 Qatar
2012 United Arab Emirates
2018 Indonesia
2012 Brazil
2009 Brazil
2006 Switzerland
2015 Brazil
2009 Spain
2009 Mexico
2003 Mexico
2012 Australia
2012 Spain
2006 Mexico
2015 Spain (Regions)
2012 Mexico
2022 Indonesia
2022 Kazakhstan
2006 Spain
2012 Italy
2006 Italy
2022 Spain
2009 Italy
2018 Canada
2006 Canada
2012 Canada
2003 Canada
2015 Canada
2018 United Kingdom
2012 Kazakhstan
2022 Uzbekistan
2012 Macao
2009 Macao
2006 Baku (Azerbaijan)


Bottom node of the split
Upper node of the split
Values of the parameters
2
0
−2
−4
−6
1
1−2
1−3
1−4
1−5
1−6
1−7
1−8
1−9
Father Occupation categories in bottom node
5.0
2.5
0.0
-2.5
-5.0
2−10
3−10
4−10
5−10
6−10
7−10
8−10
9−10
10
Father Occupation categories in upper node
Parameters
1st
2nd
3rd
4th


Values of the statistic
700
600
500
400
2−10
3−10
4−10
5−10
6−10
7−10
8−10
9−10
10
Father Occupation categories in upper node

> “We do not simply want to render the functions identical at a low level, so we need to adopt some conception of ‘maxi-minning’ these functions. […] A natural approach is therefore to maximize the area under the lower envelope of the functions.”
>
> Roemer and Trannoy (2015) p. 231


Standardized score
ECDF
Types
Blue−collar
White−collar