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).

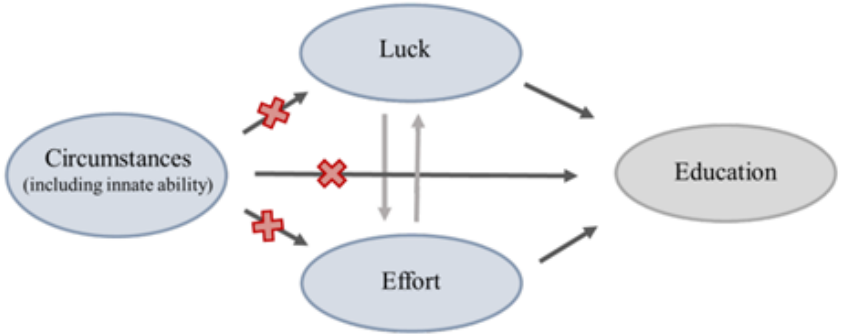
Background: IoP Measurement Approaches

(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])

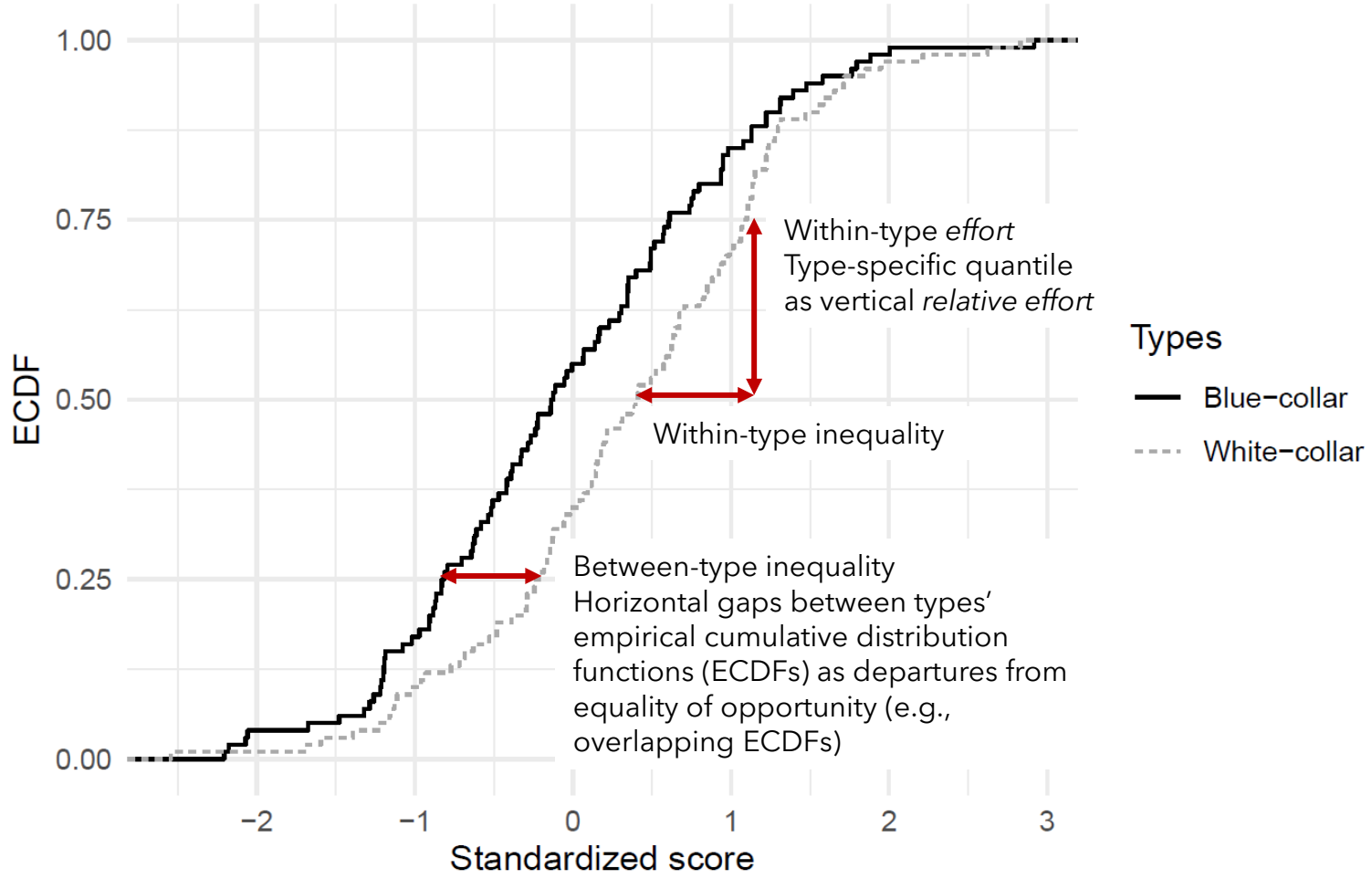
Methods: IoP Definition

"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)

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

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

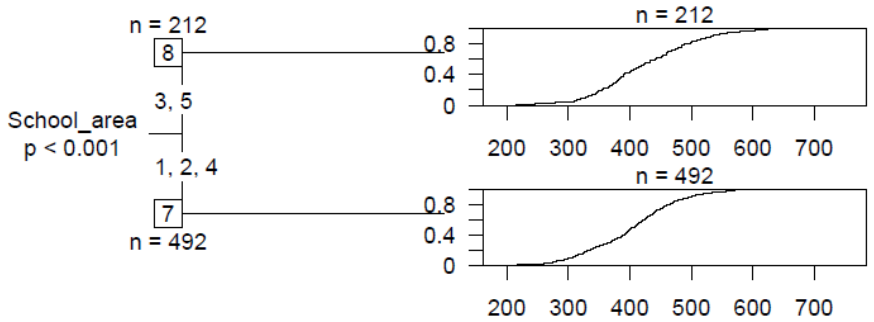
Simplified Population → 1 Circumstance & 2 Types



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)

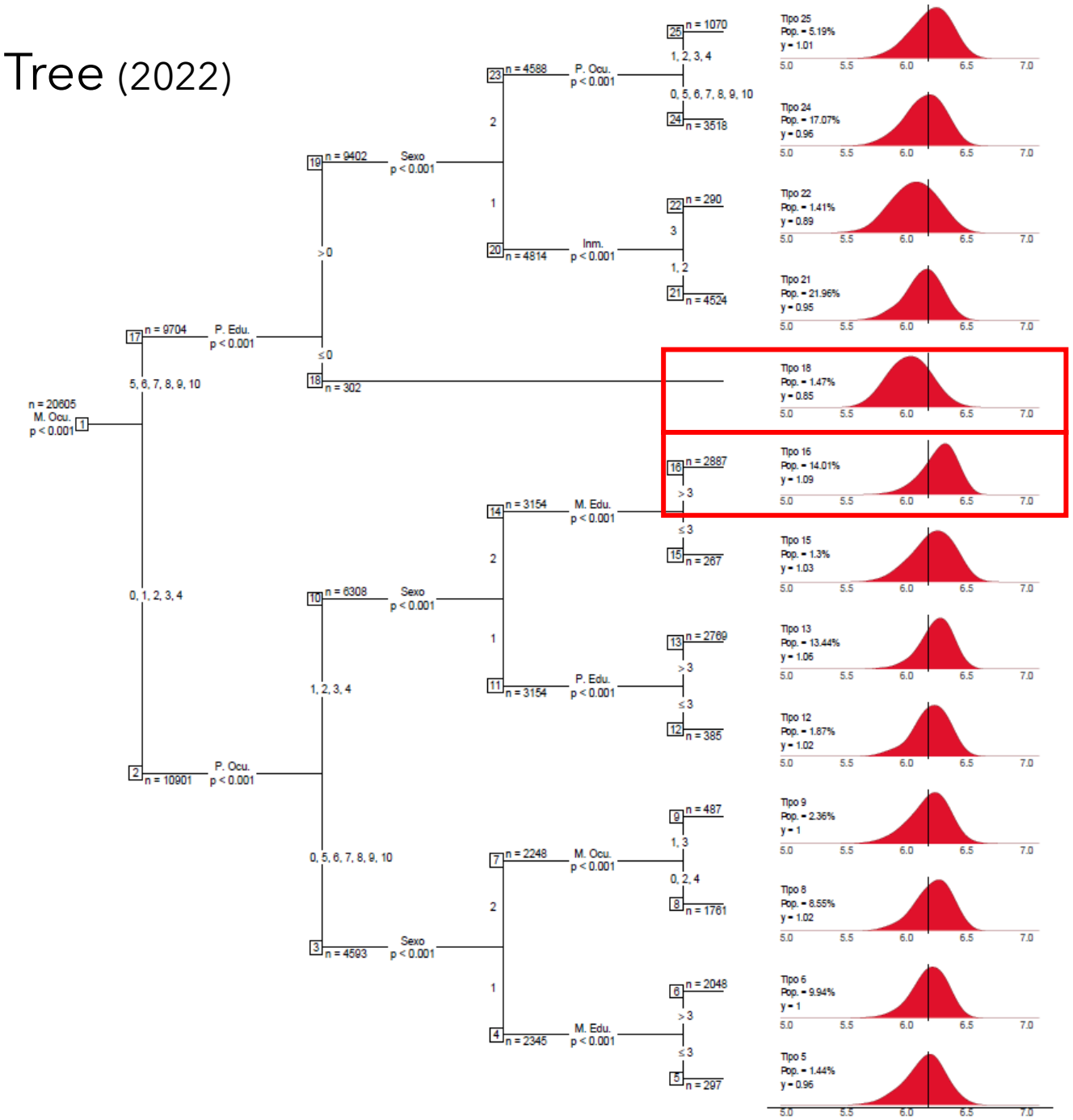
N types = 13
R^2 = IoP Variance / Total Variance = 12 %

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

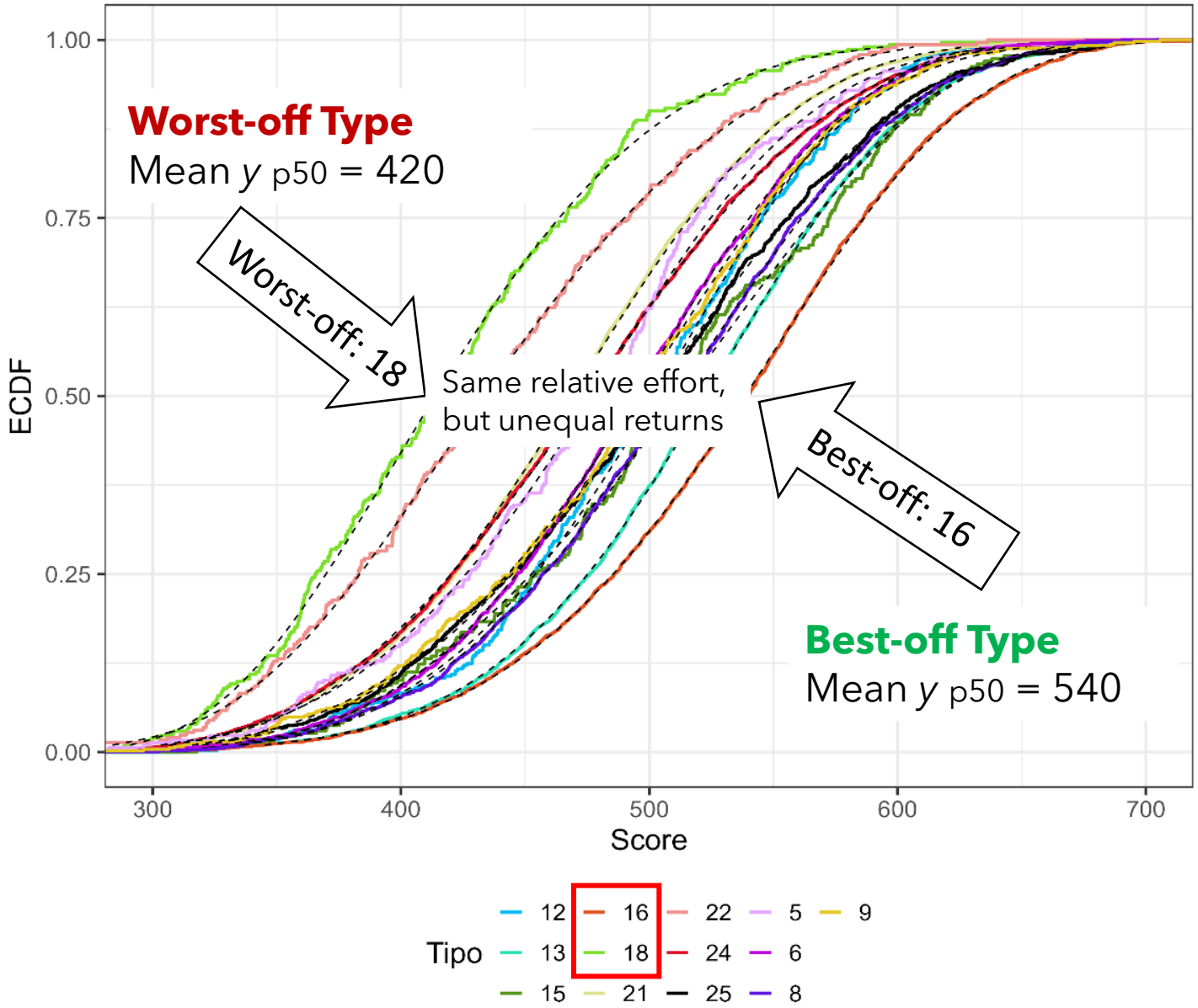
Mother Occupation: ISCO 5-10 →
Father Education: ISCED 0

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

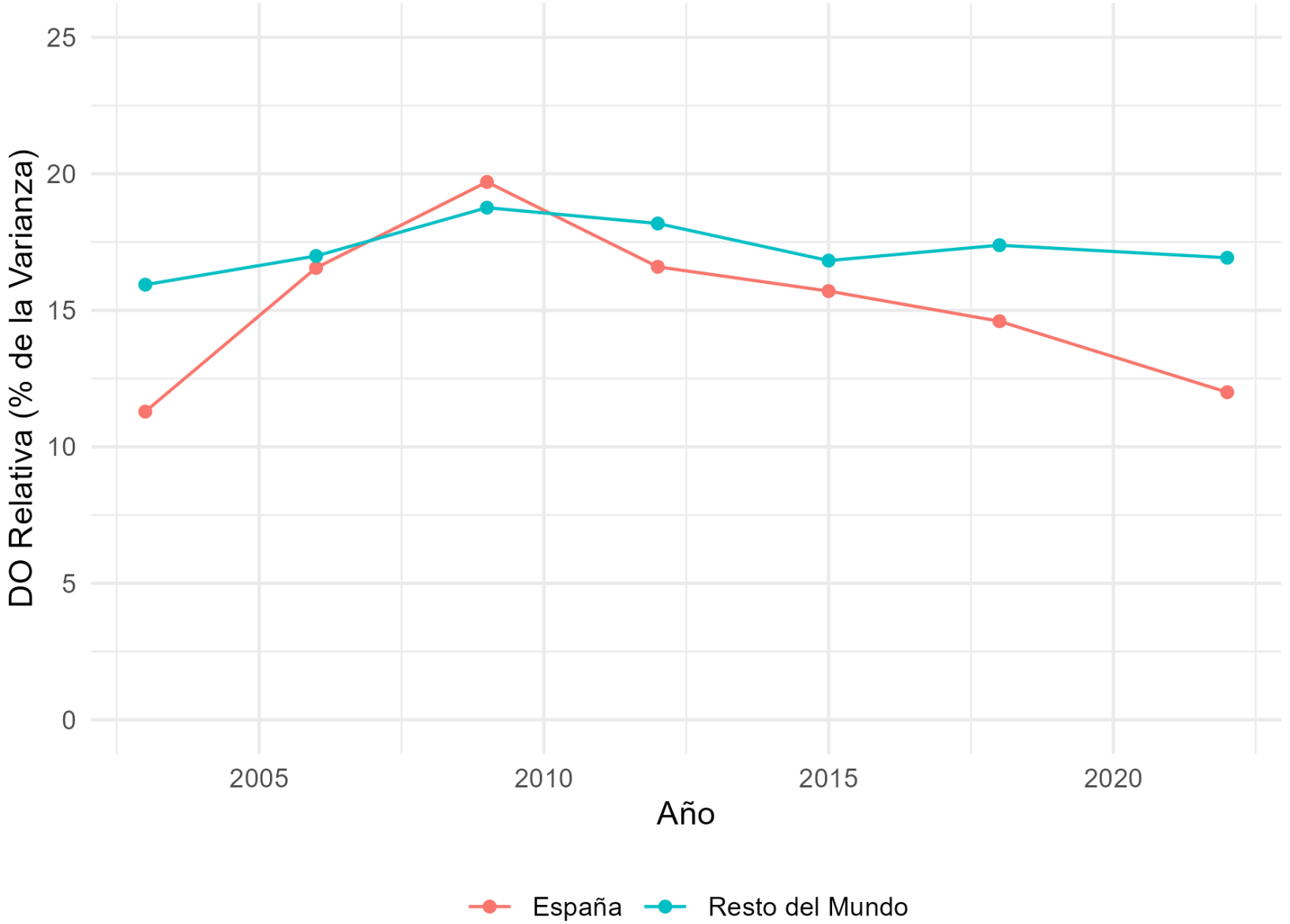
Mother Occupation: ISCO 1-4 →
Father Occupation: ISCO 1-4 →
Sex: Men →
Mother Education: ISCED > 3



Types: Empirical Cumulative Distribution Functions (2022)

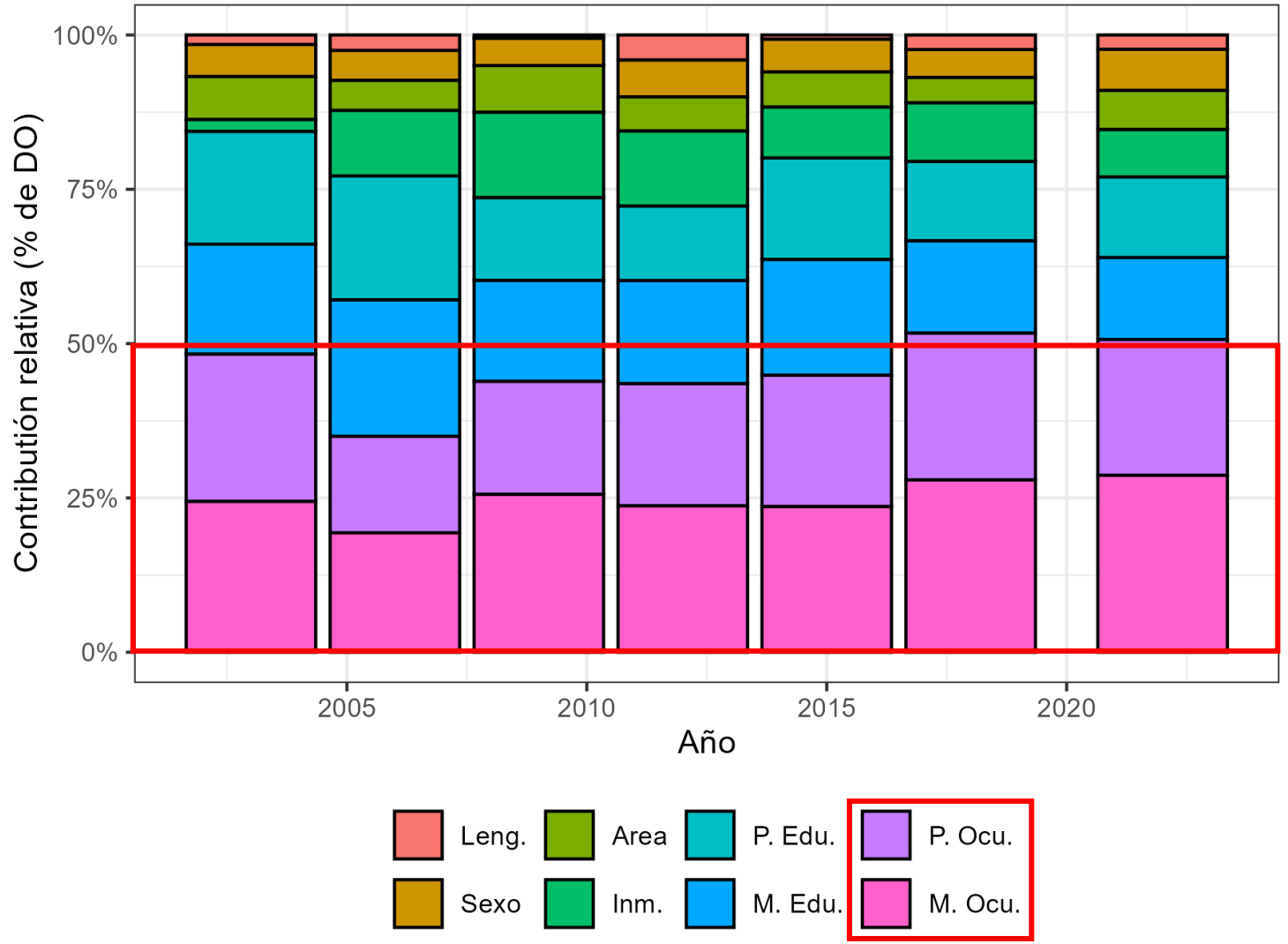


IoP Trends: % Variance



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