Model Uncertainty and Measures of Inequality of Opportunity

Giovanni Bernardo (University of Naples Federico II)

Steven Durlauf (University of Chicago)

Andros Kourtellos (University of Cyprus)

Chih Ming Tan (University of North Dakota)

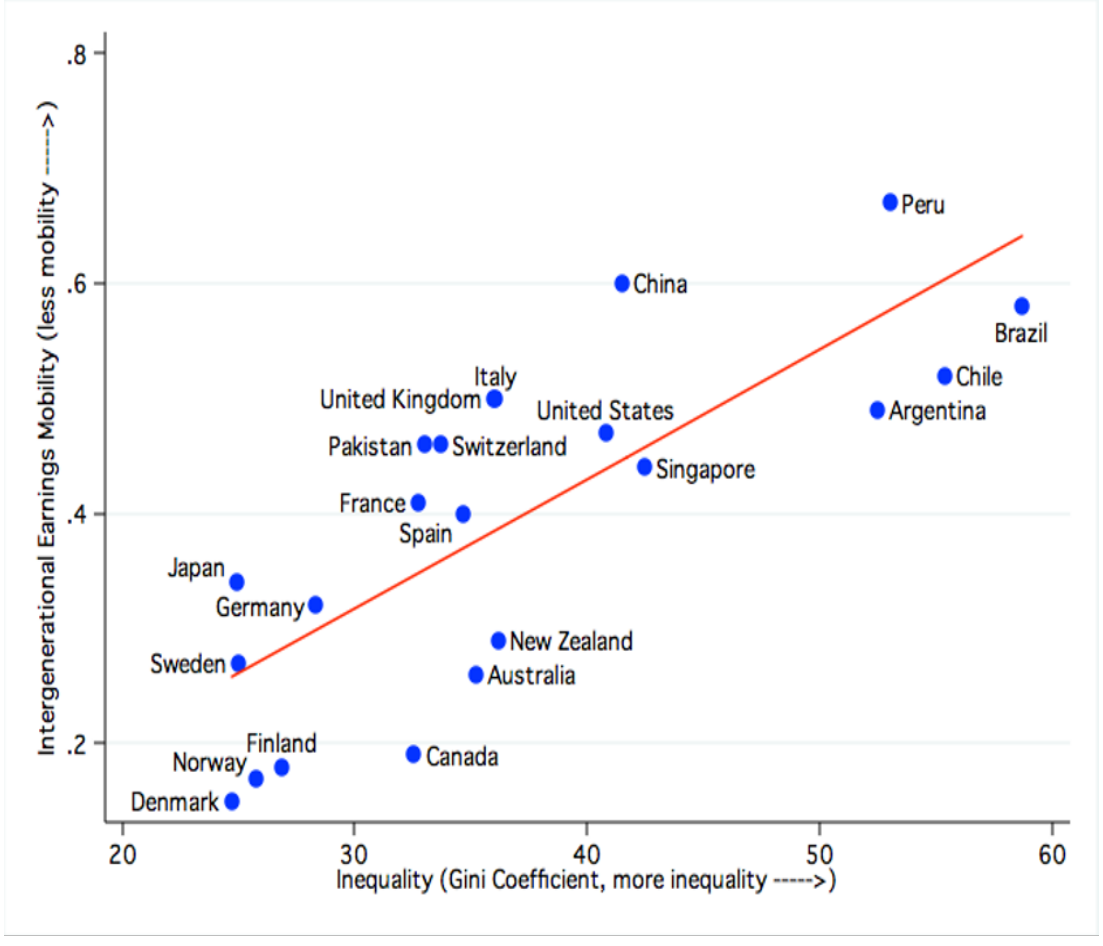
The 23rd Conference on Research on Economic Theory and Econometrics
Tinos 7-11 July 2025

INTRODUCTION

- One of the most visible stylized facts in contemporary inequality research is the association, across countries, between measures of cross-sectional income inequality and intergenerational mobility or persistence.

- This is known as the Great Gatsby Curve (GGC) (e.g., Hassler, Mora, and Zeira (2007), Andrews and Leigh (2009), and Corak (2006, 2013).

THE GREAT GATSBY CURVE



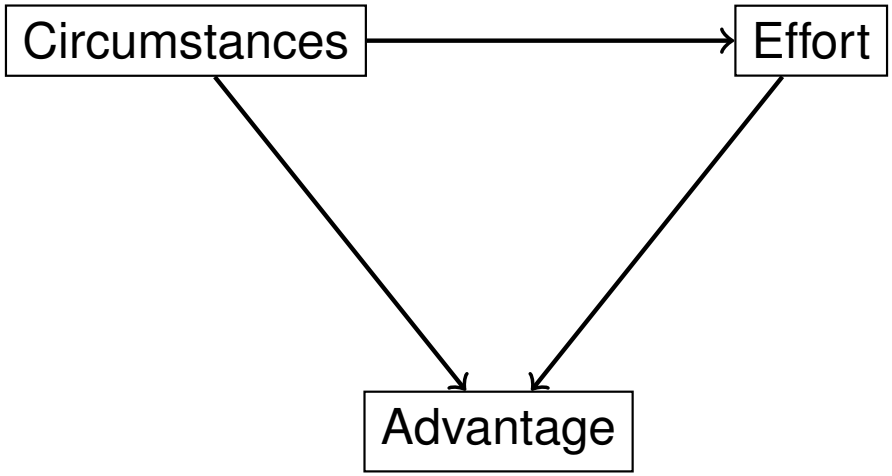
Source: Corak (2013)

- As highlighted in Durlauf, Kourtellos, and Tan (2022),
  - GGC measures of economic mobility are mechanically related to standard measures of cross-sectional earnings inequality.
  - There are five distinct classes of theories, including models on family investments, skills, social influences, political economy, and aspirations that can provide a behavioral mechanism to explain the relationship.

- How do concepts of social justice relate to GGC?

- Seminal work by Roemer (1993, 1998) established the concept of EOp as central ideal of distributive justice.

THE EOP MODEL



- Circumstances: Factors beyond people's control (e.g., gender, race, parental income)
- Effort: Choice variables for which individuals should be held responsible (e.g., number of hours worked, hours of training, occupational choice)
- Advantage: Income, Wealth, Occupation, etc.

What is the extent to which individual agency, decision-making, choices, and effort can overcome the confines of circumstances in determining individual outcomes?

- Inequality can be justified if and only if it is attributable to choices for which individuals can “reasonably” be held responsible.
- However, inequalities that derive from differences in factors that are outside of the individual’s control (circumstances) are not justified and should be mitigated by or compensated for via public policy.
- The society should do what it can to “level the playing field”.

There are two main approaches in the literature on the compensation principle.

(1) The ex-ante approach:

- Van de Gaer (1993), Lefranc, Peragine and Serlenga (2008), Pistolesi, and Trannoy (2009), Checchi and Peragine (2010), Hufe (2021).

- From the ex-ante perspective, there is equality of opportunity if differences in circumstances do not matter for an individual's opportunity set.

- The goal is to enact ex-ante policies that compensate for disadvantages that arise from such differences.

- Because the ex-ante approach assumes that compensation occurs prior to the realization of efforts, it requires equalization of some valuation of the opportunity sets available to everyone, regardless of their circumstances.

INEQUALITY OF OPPORTUNITY

- IOp is defined as the part of total inequality that can be explained by circumstances –variables beyond the individual's control such as parental income, parental education, gender, or place of birth.

- Outcome decomposition

$$Y_i = m(C_i) + \varepsilon_i, \qquad m(C_i) = \mathbb{E}[\, Y_i \mid C_i \,] = Y^*,$$

  - C_i is a p-dimensional vector of circumstance variables and
  - e_i is a scalar index variable of effort for which the individual is responsible, treated as a regression error, $E(e_i|C_i) = 0$.

- Gini-based IOp

$$I(Y^*) = \frac{1}{2}\mathbb{E}[\,|m(C_i) - m(C_j)|\,].$$

- Special Cases:
  - circumstances do not predict outcomes (No IOp) :

$$m(C_i) = \mathbb{E}[Y_i] \Rightarrow I(Y^*) = 0$$

.

  - circumstances fully determine outcomes (Maximum IOp) :

$$m(C_i) = Y_j \Rightarrow I(Y^*) = \mathbb{E}[|Y_i - Y_j|] = I(Y)$$

.

- Relative Inequality of Opportunity $RIOp = I^*(C_i)/I(Y_i)$.
- Estimation of IoP involves estimating $\widehat{m}(C_i)$ using regression or machine learning.

$$\widehat{I} = \frac{1}{2n}\sum_{i<j}\left|\widehat{m}(C_i) - \widehat{m}(C_j)\right|$$

THE CHALLENGE OF MODEL UNCERTAINTY

- Which circumstances matter?
- How to model their relationships (linear, interactions, non-linearities)?
- Different models imply different levels of predicted inequality.

THE CHALLENGE OF MODEL UNCERTAINTY

- Which circumstances matter?
- How to model their relationships (linear, interactions, non-linearities)?
- Different models imply different levels of predicted inequality.

SOLUTION

Use complementary strategies to address uncertainty:

- Linear models: e.g., KS, BIC, Linear-BMA.
- Machine learning algorithms: BART-BMA, C-Forest, XGBoost.

ESTIMATION METHODS FOR IOP: LINEAR-BMA

- **Linear-BMA:**
  - Estimates are based on a model space of all subsets of circumstances.
  - Integrates model uncertainty via posterior-weighted model averaging.
  - Produces robust estimates by avoiding reliance on a single model.
  - Assumes linear and additive relationships; may miss nonlinearities or interactions.

ESTIMATION METHODS FOR IOP: RF AND C-FOREST

- **Random Forests (RF):**
  - Builds an ensemble of bootstrapped decision trees using CART splits.
  - Captures nonlinearities and interactions via recursive partitioning.
  - May be biased toward variables with more categories or higher variance.
  - Averages predictions to reduce variance (bagging).
- **Conditional Inference Forests (C-Forest):**
  - Uses permutation-based tests for unbiased variable selection.
  - Prevents overfitting by stopping at statistically insignificant splits.
  - Handles mixed-type variables well and reduces RF's selection bias.
  - Offers more interpretable and statistically sound trees.

ESTIMATION METHODS FOR IOP: BART-BMA

- **BART-BMA:**
  - Builds models as sums of regression trees using greedy search.
  - Incorporates model uncertainty via Bayesian Model Averaging.
  - Captures complex nonlinearities and interactions with prior-based regularization.
  - Avoids MCMC by relying on efficient model selection and averaging.

ESTIMATION METHODS FOR IOP: XGBOOST

- **XGBoost:**
  - Implements gradient boosting with regularization and shrinkage.
  - Sequentially adds trees to correct residual errors and reduce bias.
  - Highly scalable and efficient; popular in practical machine learning tasks.
  - Relies on boosting rather than Bayesian model averaging to combine models.

DATA

- We employ the European Union Statistics on Income and Living Conditions (EU-SILC; 2011 and 2019).
- Launched in 2003 with aim of collecting timely and comparable cross-sectional and longitudinal data on income, poverty, social exclusion and living conditions in European countries.
- For each country, took a random sample of all resident private households.
- The 2011 wave covers a cross-section of 31 European countries. All current 27 EU members plus UK, Switzerland, Iceland, and Norway.
- The 2019 wave covers a cross-section of 29 European countries. All current 27 EU members plus Switzerland and Norway.

Dependent variable: Logarithm of equivalized disposable household income as a proxy for the child's income, adjusted for the square root of household size.

Circumstances: We use three distinct sets of circumstances pertaining to the households of the respondents at the age of 14.

(I) Father characteristics, including country of origin, citizenship, educational level, occupational status, and type of occupation;

(II) Mother characteristics, including country of origin, citizenship, educational level, and occupational status; and

(III) Household characteristics, including tenancy status, household financial situation, ability to meet end-of-month expenses (not included in 2019), family type, number of adults and children, and number of workers.

- We have restricted our sample to include only respondents aged between 30 and 60.
- Non-ordered categories - such as employment status or tenancy status -were reconfigured into dummy variables. Conversely, ordinal variables, exemplified by education level or household financial situation, were retained in their original format.
- We grouped employment types into three macro-categories: high-skilled non-manual, low-skilled non-manual, and skilled manual and elementary occupations. The unemployed, unknown, and dead were grouped separately.

INCOME DATA (2011 & 2019, PART 1)

TABLE: Summary statistics by country. Disposable income PPP-adjusted.

Country	Obs.	Mean	Std. Dev.	Gini	Obs.	Mean	Std. Dev.	Gini
AT	6220	25822.59	13559.18	0.267	5398	25772.31	13329.88	0.269
BE	6011	21989.00	10210.00	0.248	6212	23174.09	10385.38	0.235
BG	7146	8044.97	5379.53	0.331	6715	11702.34	10339.40	0.391
CH	7583	26536.39	14853.23	0.278	7096	31667.28	17161.06	0.282
CY	4589	22296.34	12128.00	0.278	4362	22586.62	14208.04	0.286
CZ	8711	13701.82	6695.02	0.253	7283	16688.85	7077.73	0.228
DE	12683	24018.78	12735.05	0.285	9732	26010.14	13949.47	0.273
DK	5795	25743.47	10301.00	0.228	4367	28372.49	11811.74	0.241
EE	5338	10355.61	6288.50	0.330	6022	15632.68	8002.69	0.288
EL	6184	12864.69	8905.98	0.334	14733	11107.59	6858.92	0.311
ES	15481	18084.27	11094.19	0.328	17229	18368.74	11090.94	0.324
FI	9743	24533.47	11940.67	0.243	9454	26122.18	12763.45	0.245
FR	11078	23251.59	13798.00	0.287	10255	23601.26	12914.76	0.268
HR	6945	9828.00	5764.00	0.306	7515	12268.35	6634.01	0.281
HU	13330	9988.97	5456.16	0.275	5574	9593.48	4893.29	0.277
IE	4318	20704.10	11970.00	0.295	4238	22999.76	12703.41	0.269

INCOME DATA (2011 & 2019, PART 2)

TABLE: Summary statistics by country. Disposable income PPP-adjusted.

Country	Obs.	Mean	Std. Dev.	Gini	Obs.	Mean	Std. Dev.	Gini
IS	3682	22628.31	9242.70	0.221	N/A	N/A	N/A	N/A
IT	21070	20420.00	12263.00	0.314	17613	20470.65	11663.72	0.317
LT	5403	8786.89	5604.43	0.344	4801	15070.94	10041.63	0.343
LU	6765	28799.00	15451.00	0.270	4719	28842.45	18281.41	0.317
LV	6423	8188.93	5747.00	0.362	4509	12976.28	8622.96	0.336
MT	4431	17155.64	8950.54	0.271	3538	20727.47	10759.52	0.270
NL	11411	24354.62	10672.32	0.245	11362	24563.26	11999.04	0.255
NO	5026	28171.47	10868.00	0.206	6207	28909.88	12922.14	0.231
PL	15545	11235.09	7128.00	0.316	19743	13957.88	7991.95	0.288
PT	5899	12731.00	8849.00	0.333	13767	13198.72	8683.95	0.303
RO	7820	5636.55	3530.00	0.336	7217	9284.36	5653.28	0.342
RS	N/A	N/A	N/A	N/A	6492	6741.84	4308.51	0.329
SE	6599	20500.91	8342.00	0.237	5206	22387.92	9585.13	0.253
SI	13183	10796.54	4630.00	0.234	10756	12060.45	5175.58	0.230
SK	6779	12256.34	5624.33	0.256	6073	11156.28	4500.36	0.224
UK	7391	21572.09	14182.21	0.320	N/A	N/A	N/A	N/A

Performance Assessment

Goal: Assess how well each model predicts income out-of-sample.

- We use **10-fold cross-validation**.
  - Dataset split into 10 subsets (folds)
  - Each fold used once as test set; remaining as training set
  - Predict log-income on test fold and compute squared error
- **Out-of-sample Mean Squared Error (MSE^{OOS})**:

$$\text{MSE}^{OOS} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}(Y_i - \widehat{Y}_i^k)^2$$

- We evaluate MSE^{OOS} on **log-income** (interpretable as percentage error).

OUT-OF-SAMPLE MSE (PART 1)

TABLE: Mean Squared Error for Test Set (2011 and 2019)

	2011				2019			
Country	Linear-BMA	BART-BMA	C-Forest	XGBoost	Linear-BMA	BART-BMA	C-Forest	XGBoost
AT	0.25	0.253	0.249	0.249	0.316	0.303	0.299	0.307
BE	0.3	0.305	0.297	0.302	0.3	0.3	0.294	0.303
BG	0.347	0.365	0.334	0.33	0.37	0.386	0.372	0.365
CH	0.347	0.352	0.346	0.352	0.285	0.288	0.282	0.284
CY	0.219	0.225	0.219	0.219	0.247	0.251	0.248	0.249
CZ	0.198	0.202	0.201	0.202	0.151	0.151	0.15	0.15
DE	0.484	0.486	0.481	0.489	0.538	0.539	0.537	0.545
DK	0.792	0.748	0.747	0.792	0.656	0.665	0.649	0.687
EE	0.586	0.591	0.583	0.587	0.722	0.728	0.721	0.739
EL	1.371	1.388	1.366	1.383	0.867	0.806	0.792	0.805
ES	1.08	1.087	1.08	1.088	0.734	0.741	0.733	0.734
FI	0.371	0.376	0.367	0.385	0.289	0.301	0.3	0.305
FR	0.311	0.316	0.309	0.31	0.262	0.263	0.261	0.261
HR	0.924	0.929	0.916	0.933	0.464	0.463	0.459	0.46
HU	0.238	0.248	0.237	0.236	0.407	0.402	0.391	0.397
IE	0.475	0.444	0.439	0.44	0.206	0.206	0.204	0.203

OUT-OF-SAMPLE MSE (PART 2)

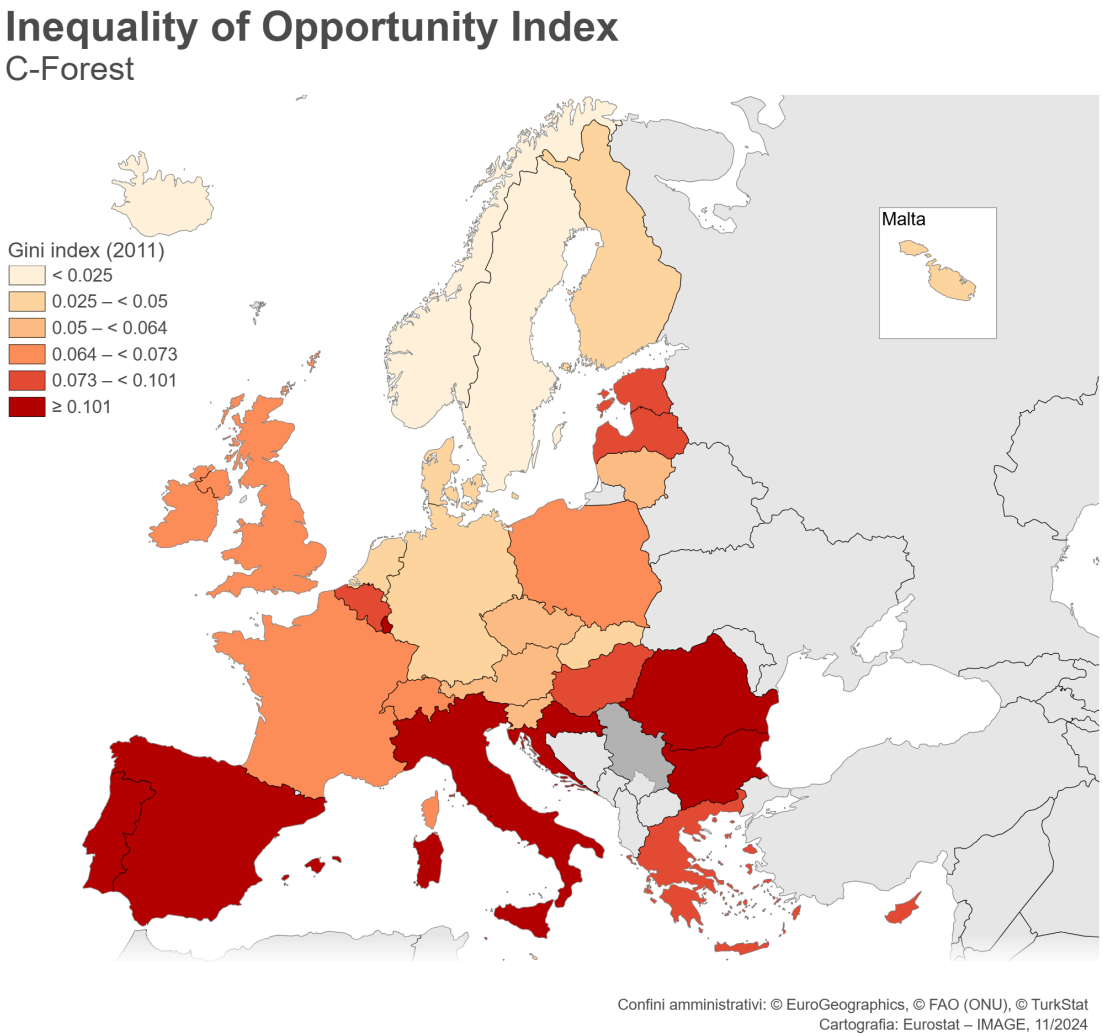
TABLE: Mean Squared Error for Test Set (2011 and 2019)

	2011				2019			
Country	Linear-BMA	BART-BMA	C-Forest	XGBoost	Linear-BMA	BART-BMA	C-Forest	XGBoost
IS	0.444	0.462	0.442	0.48	NA	NA	NA	NA
IT	1.056	1.066	1.05	1.052	0.951	0.955	0.949	0.958
LT	0.886	0.914	0.9	0.911	0.677	0.683	0.671	0.682
LU	0.51	0.522	0.507	0.517	0.472	0.475	0.465	0.469
LV	0.975	0.996	0.967	0.973	0.896	0.907	0.879	0.897
MT	0.344	0.351	0.344	0.348	0.289	0.302	0.298	0.302
NL	0.357	0.382	0.378	0.388	0.646	0.563	0.546	0.585
NO	0.231	0.246	0.242	0.254	0.367	0.37	0.363	0.383
PL	0.383	0.391	0.381	0.382	0.58	0.583	0.58	0.581
PT	0.329	0.341	0.329	0.328	0.329	0.338	0.328	0.326
RO	0.467	0.48	0.462	0.463	0.846	0.927	0.913	0.919
SE	0.61	0.579	0.552	0.581	0.923	0.796	0.747	0.842
SI	0.181	0.181	0.18	0.18	0.171	0.172	0.17	0.171
SK	0.217	0.219	0.217	0.219	0.256	0.238	0.231	0.231
UK	0.987	0.991	0.984	0.999	NA	NA	NA	NA
RS	NA	NA	NA	NA	0.968	0.988	0.961	0.966

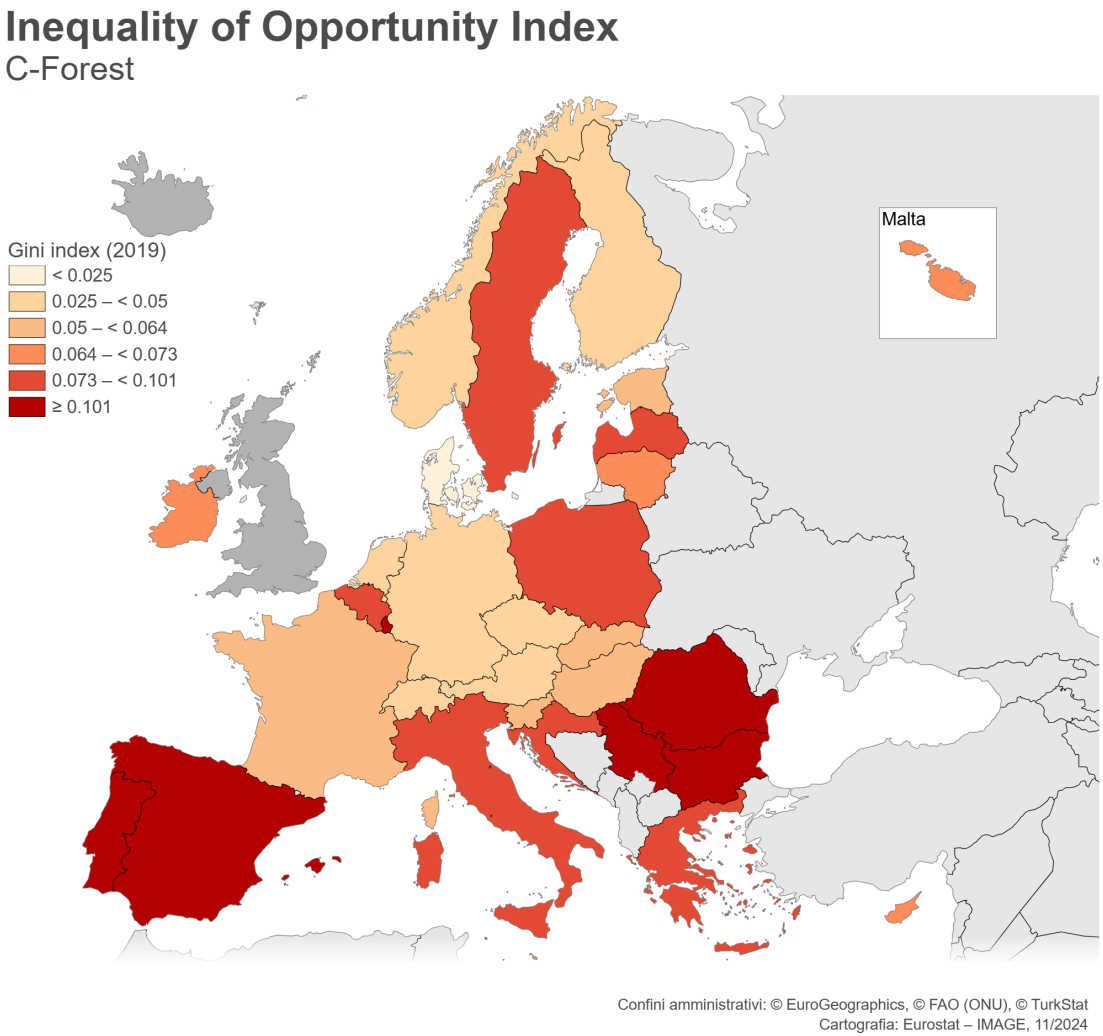
MODEL COMPARISON: PREDICTIVE ACCURACY (MSE)

- **C-Forest** consistently achieves the lowest MSE^{OOS}, reflecting strong performance in capturing nonlinearities and interactions.
- **XGBoost and Linear-BMA** also perform well, though slightly less robust than C-Forest.
- **BART-BMA** often exhibits the highest MSE^{OOS}, possibly due to overfitting or poor adaptation to this data context.

INEQUALITY OF OPPORTUNITY ACROSS EUROPEAN COUNTRIES (WAVE 2011)



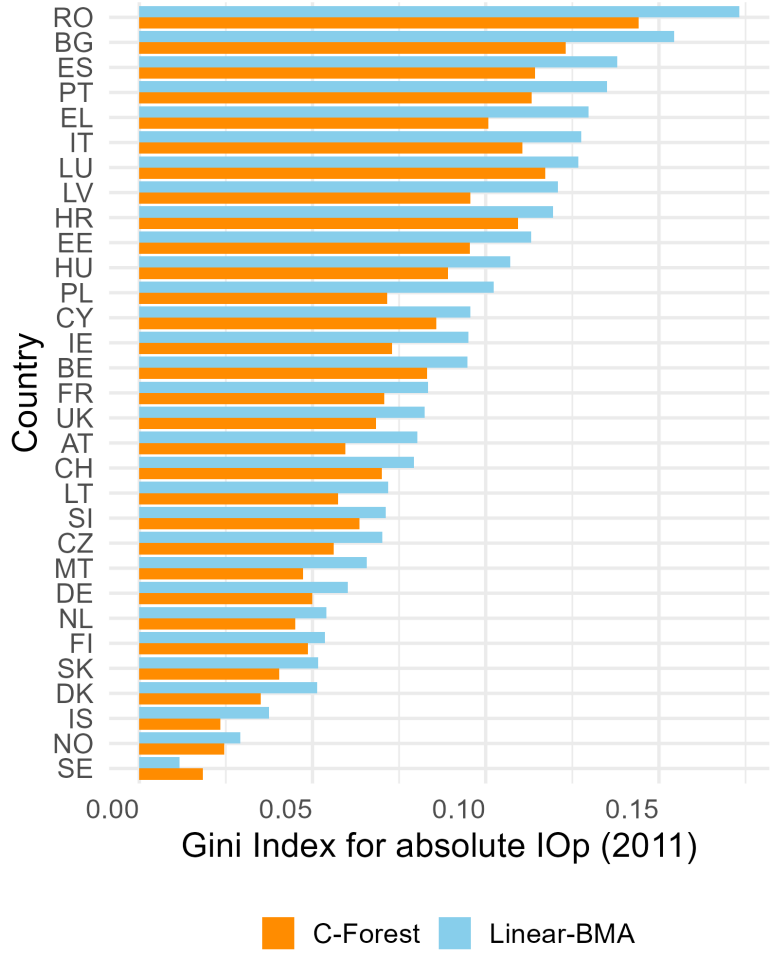
INEQUALITY OF OPPORTUNITY ACROSS EUROPEAN COUNTRIES (WAVE 2019)

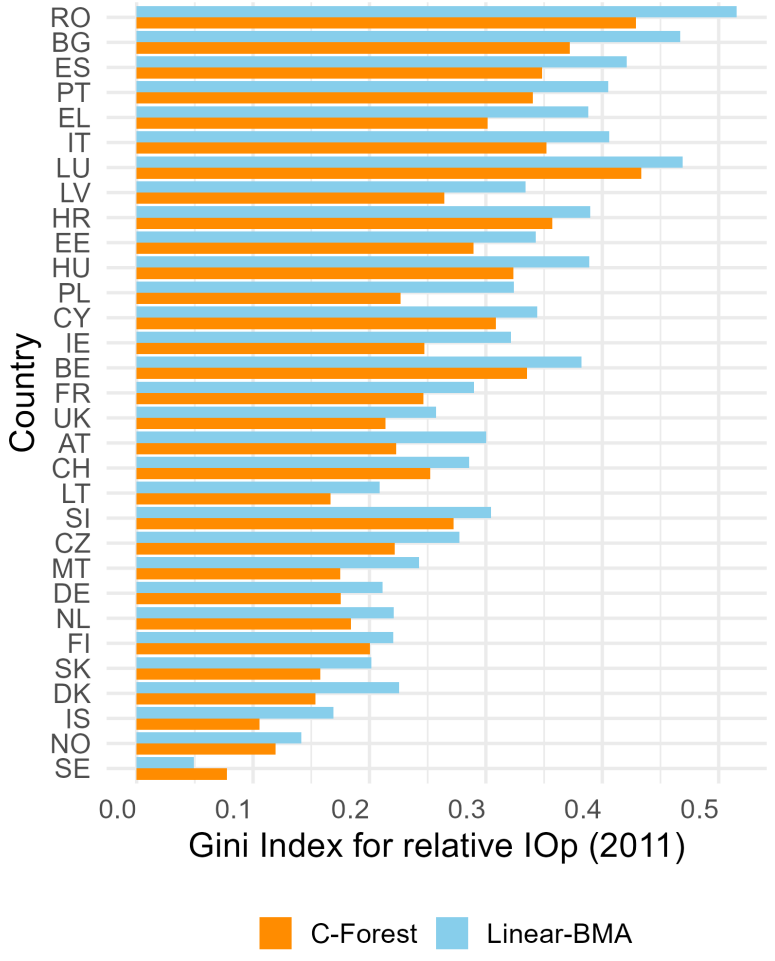


SWEDISH EXCEPTIONALISM (2011–2019)

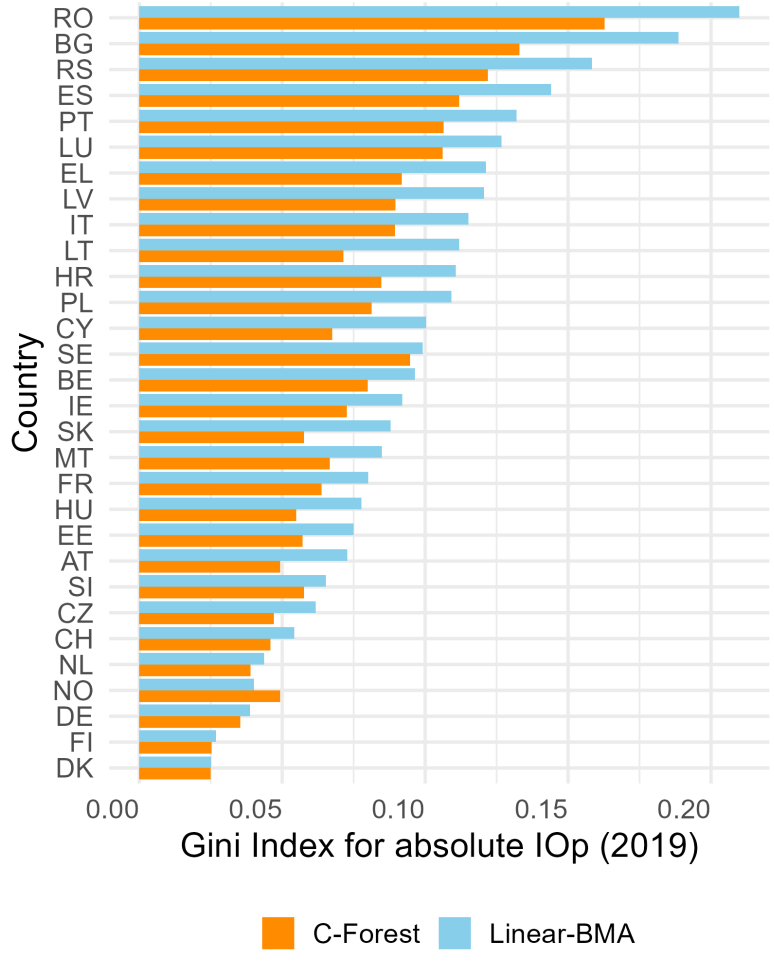
- Sweden exhibits the most pronounced increase in inequality of opportunity (IOp) over the period.
- 2011 estimates affected by missing data, but still consistent with low-IOp Nordic patterns.
- By 2019, Sweden transitions from a top-performing country to a mid-level position in the IOp distribution (just below Cyprus).
- While income inequality may have stabilized post-2011, IOp continued to deteriorate.
- In 2019, parental birthplace (non-EU/Europe) emerges as the dominant predictor of IOp, in contrast to its negligible role in 2011.
- This shift highlights the growing relevance of migration-related factors in shaping opportunity structures.

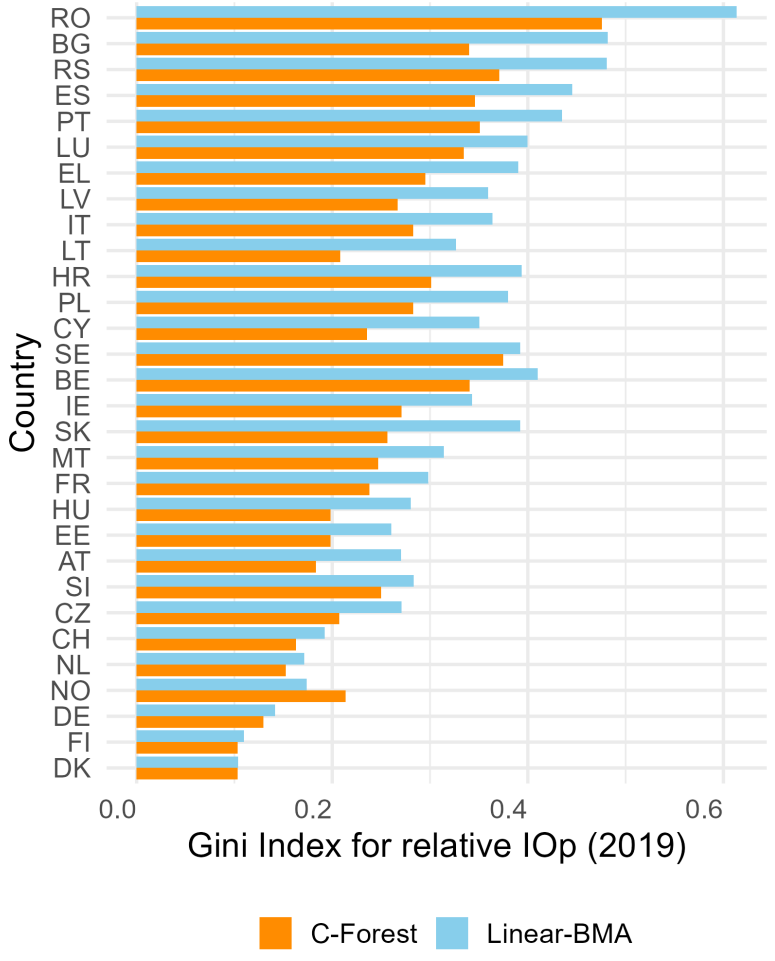
IOP ESTIMATES – LINEAR-BMA AND C-FOREST (2011)





IOP ESTIMATES – LINEAR-BMA AND C-FOREST (2019)





TOP CIRCUMSTANCES – (PART 1)

TABLE: LINEAR-BMA (PIP, Post. Mean, Share of Countries)

	2011						2019					
Variable	Post. mean	Avg. PIP	>50%	>75%	>95%	>99%	Post. mean	Avg. PIP	>50%	>75%	>95%	>99%
Education (F)	0.09	0.81	0.81	0.71	0.61	0.58	0.08	0.72	0.70	0.70	0.57	0.50
Education (M)	0.07	0.63	0.65	0.58	0.58	0.52	0.05	0.53	0.53	0.47	0.37	0.27
Born Non-EU (I)	-0.11	0.40	0.42	0.26	0.26	0.23	-0.13	0.41	0.40	0.37	0.33	0.33
Children (H)	-0.02	0.43	0.42	0.35	0.29	0.29	-0.02	0.43	0.43	0.37	0.33	0.30
Skilled occ. (F)	-0.04	0.40	0.35	0.29	0.26	0.23	-0.05	0.44	0.40	0.37	0.23	0.17
Financial sit. (H)	0.01	0.29	0.26	0.23	0.19	0.13	0.03	0.49	0.47	0.47	0.40	0.37

TABLE: BART-BMA (PIP and Share of Countries)

	2011					2019				
Variable	Avg. PIP	>50%	>75%	>95%	>99%	Avg. PIP	>50%	>75%	>95%	>99%
Adults (H)	0.50	0.71	0.42	0.42	0.42	0.52	0.77	0.57	0.33	0.33
Children (H)	0.31	0.68	0.19	0.19	0.13	0.42	0.60	0.40	0.30	0.27
Workers (H)	0.42	0.61	0.35	0.35	0.29	0.24	0.47	0.20	0.13	0.13
Financial sit. (H)	0.22	0.52	0.16	0.10	0.06	0.39	0.57	0.37	0.30	0.23
Education (F)	0.17	0.48	0.06	0.03	0.03	0.10	0.37	0.00	0.00	0.00

TOP CIRCUMSTANCES – (PART 2)

TABLE: C-Forest (Variable Importance and Share of Countries)

	2011					2019				
Variable	VarImp	>50%	>75%	>95%	>99%	VarImp	>50%	>75%	>95%	>99%
Education (F)	0.73	0.81	0.61	0.39	0.39	0.65	0.70	0.57	0.40	0.37
Education (M)	0.54	0.55	0.45	0.32	0.29	0.43	0.43	0.27	0.13	0.10
Skilled occ. (F)	0.41	0.35	0.13	0.10	0.06	0.35	0.20	0.10	0.00	0.00
Born Non-EU (I)	0.30	0.29	0.13	0.06	0.06	0.38	0.40	0.27	0.17	0.17
High-skill occ. (F)	0.28	0.23	0.10	0.06	0.06	0.25	0.27	0.10	0.10	0.10
Born outside EU (F)	0.11	0.10	0.03	0.00	0.00	0.22	0.23	0.13	0.07	0.03

TABLE: XGBoost (Variable Importance and Share of Countries)

	2011					2019				
Variable	VarImp	>50%	>75%	>95%	>99%	VarImp	>50%	>75%	>95%	>99%
Financial sit. (F)	0.61	1.00	0.55	0.35	0.19	0.71	1.00	0.73	0.47	0.30
Education (F)	0.69	0.97	0.74	0.45	0.19	0.59	1.00	0.57	0.30	0.17
Children (H)	0.60	0.97	0.55	0.35	0.19	0.69	1.00	0.70	0.47	0.27
Adults (H)	0.49	0.94	0.48	0.16	0.10	0.45	0.90	0.43	0.17	0.03
Workers (H)	0.47	0.90	0.45	0.19	0.10	0.35	0.83	0.33	0.07	0.07
Education (M)	0.53	0.84	0.55	0.29	0.19	0.45	0.87	0.37	0.27	0.13

GREAT GATSBY CURVE

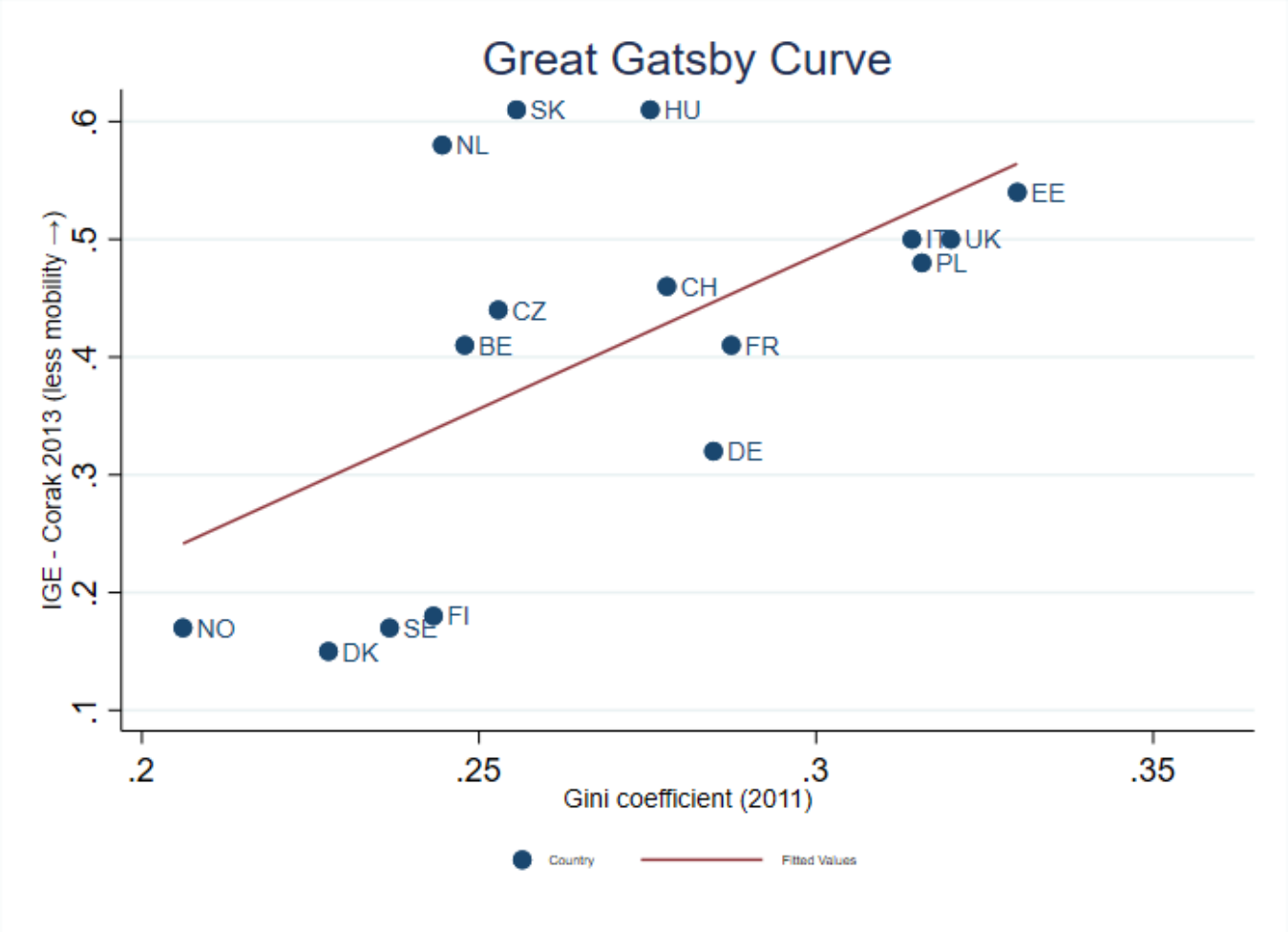
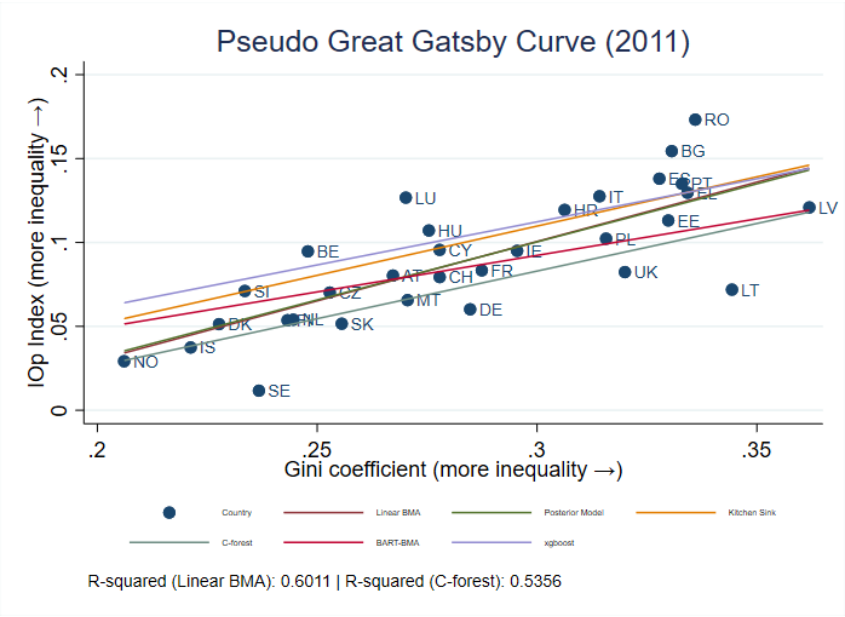
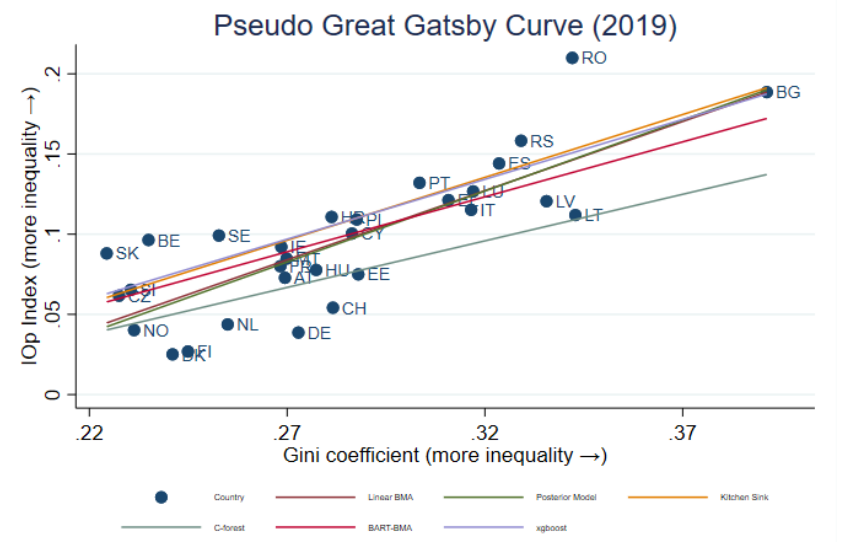


FIGURE: This figure shows the relationship between Intergenerational Elasticity of Income and Income Inequality.

PSEUDO GREAT GATSBY CURVES (2011 AND 2019)



R-squared (Linear BMA): 0.6011 | R-squared (C-forest): 0.5356



R-squared (Linear BMA): 0.6409 | R-squared (C-forest): 0.5459

CONCLUSION

- We examined IOp in 31 EU countries (2011–2019) using Linear-BMA and C-Forest, alongside BART-BMA and XGBoost.
- C-Forest performs best among tree-based methods. Linear-BMA also outperforms the traditional "kitchen sink" model by addressing model uncertainty and avoiding overestimation.
- Linear-BMA and C-forest analysis initially identifies key factors such as parental education levels, father's occupation and the respondent's citizenship status as consistently drive IOp across country
- XGBOOST and BART-BMA revealed additional robust circumstances: specifically, the critical role played by the respondent's family of origin at age 14, including financial conditions, the number of working household members, and the number of children in the household.
- This variation underscores the absence of a one-size-fits-all model, emphasizing the necessity of accounting for model uncertainty in IOp estimation.
- Accounting for model uncertainty and heterogeneity is essential for robust IOp estimation. Advanced methods provide better guidance for both research and policy.

Thank you for your attention!

Questions and comments are welcome.

Giovanni Bernardo
giovanni.bernardo@unina.it